Bayesian Fusion in Cancer Gene Prediction

J Das
Institute of Radio Physics & Electronics, University of Calcutta,
92, APC Road, Kolkata-700009, India

S Barman
Institute of Radio Physics & Electronics, University of Calcutta,
92, APC Road, Kolkata-700009, India

ABSTRACT
Diverse high throughput genomic data is available in public domain. However, no single source data analysis technique is available even today which can fully reveal the function of genes. Therefore fusion of multiple data source using Bayesian algorithm is proposed here for prediction of genes. Amino acids sequence of proastate, colon, breast, gastric genes from National Health Informatics site are taken as source data for prediction. The spectrum of genes is fused successfully using Bayesian algorithm to screen out cancer gene from healthy gene and validated the approach with the existing DSP based prediction method.

Keywords
DFT, Bayesian Fusion Technique, Cancer Gene, Disease Diagnosis, Amino acid

1. INTRODUCTION
DNA sequence analysis is one of the major research areas of Genomic Signal Processing. DNA contains the genetic information of living organisms [1]. DNA sequence can be divided into genes and inter-genic spaces. A gene can again be divided into two sub-regions named Exon (coding region) and Intron (non-coding region). The nucleotide bases (A, T, C and G) in exon region are divide into three adjacent bases called codon which is translated into amino acid. 64 possible combinations of codons generate 20 amino acids (Table 1) [2]. Protein, regarded as a sequence of amino acids is another important part of life which drives the biological processes of living organism [3]. Deficiency of amino acids in human body may leads to different types of genetic abnormalities [4]. Any kind of abnormality present in the DNA sequence is responsible for different types of genetic diseases, like cancer. Therefore prediction of cancer gene, specially using signal processing techniques has become a challenge to the researchers. Several new ideas have been employed so far [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Such as DFT power spectrum along with digital IIR filtering is applied on BRCA1 , BRCA2 genes for their identification [3]. Digital signal processing and multivariate analysis technique PCA is applied on prostate genes to differentiate cancer and healthy genes[4]. Similarly, S.Barman (Mandal) and et. al. have used spectral estimation followed by an LPF for the identification of cancerous genes [13]. According to medical researchers some amino acids play a crucial rolein cancer diseases. Bayesian Fusion basically is a statistical method and often applied for finding the randomness in sensory data. Multisensory data fusion is most common algorithm in probabilistic robotics [15, 16, 17, 18]. In order to distinguish the cancer gene from healthy gene accurately, a novel approach, the statistical Bayesian Fusion Algorithm along with DSP technique has been explored in the present paper. Bayesian Fusion Framework is used to solve the problem of source diversity. It integrates multiple heterogeneous data sources to characterize the gene effectively. The algorithm is tested on several healthy and cancer genes of prostate, breast, colon and gastric cells, which are collected from NCBI genbank [19]. The paper is organized as follows: Introduction, Methodology, Comparison with existing method, Result and Discussion and Conclusion.

2. METHODOLOGY
2.1 Representation of Amino Acid Sequences using EIIP Mapping Technique

Table1. List of twenty amino acids and codons with their corresponding EIIP values

Amino Acids	Abbreviation		Codons	EIIP Values
Alanine	A	Ala	GCA,GCC,GCG,GCT	0.037
Cystein	C	Cys	TGC,TGT	0.082
Aspartic Acid	D	Asp	GAC,GAT	0.126
Glutamic Acid	E	Glu	GAA,GAG	0.005
Phenylalanine	F	Phe	TTC,TTT	0.094
Glycine	G	Gly	GGA,GGC.GGG,GGT	0.005
Histidine	H	His	CAC,CAT	0.024
Isoleucine	I	Ile	ATA,ATC,ATT	0.000
Lysine	K	Lys	AAA,AAG	0.037
Leucine	L	Leu	TTA,TTG,CTA,CTC, CTG,CTT	0.000
Methionine	M	Met	ATG	0.082
Asparagine	N	Asn	AAC,AAT	0.003
Proline	P	Pro	CCA,CCC,CCG,CCT	0.019
Glutamine	Q	Gln	CAA,CAG	0.076
Arginine	R	Arg	AGA,AGG,CGA,CGC ,CGG,CGT	0.095
Serine	S	Ser	AGC,AGT,TCA,TCC, TCG,TCT	0.082
Threonine	T	Thr	ACA,ACC,ACG,ACT	0.094
Valine	V	Val	GTA,GTC,GTG,GTT	0.005
Tryptophan	W	Trp	TGG	0.054
Tyrosine	Y	Tyr	TAC,TAT	0.051

Genomic data is discrete sequence of alphabets, available in the public domain. DSP has been proven as an effective tool to process this data after conversion into numerical sequence. Hence a well-known single sequence electron ion interaction pseudo potential (EIIP) numerical rule [20], based on the distribution of free electron's energy along DNA, is applied

here on genomic data. Twenty amino acids and codons with their corresponding EIIP values are listed in Table 1.

Suppose, an amino acid chain of a gene consists of

$x(n)$ = [M S A R V R S R S R G R G D]

After EIIP mapping the sequence become

$x(n)$ = [0.0823 0.0829 0.0373 0.0959 0.0057 0.0959 0.0829 0.0959 0.0829 0.0959 0.0050 0.0959 0.0050 0.1263].

After this conversion the sequence is ready to process using DSP. This type of mapping technique is used in this paper to convert the healthy and cancer prostate, breast, colon and gastric genes.

2.1 Bayesian Fusion Technique

In order to predict cancer gene more precisely, a novel approach has been considered in the paper. Bayesian Fusion is used as a predictor because of its several advantages. It allows to combine highly dissimilar types of data, converting them to a common probabilistic framework without unneccessary simplification. As genomic data are unlike in nature, this method will be well suited for gene prediction. The Bayesian Fusion Technology is applied on the spectrum of genes to successfully identify cancer and healthy genes. The method of Cancer Gene Prediction is depicted in Fig. 1 and the algorithm is as follows:

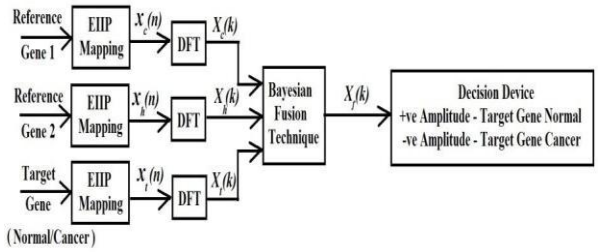

Fig. 1 Block diagram representation of Cancer Gene Prediction Technique

Algorithm: Suppose, one healthy gene (AF331165.1) and one cancer gene (NP001035756.1) have been chosen as a reference.

Step-1: Scan $x_h(n)$ is the healthy and $x_c(n)$ is the cancer reference gene respectively. Scan target gene $x_t(n)$ which may be cancer or healthy.

Step-2: Obtain DFT of $x_h(n)$, $x_c(n)$ and $x_t(n)$ as $X_h(k)$, $X_c(k)$ and $X_t(k)$ respectively using equation (1).

$$X[k] = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N} \quad (1)$$

where, $n = 0,1,\ldots\ldots, N-1$ and $k = 0,1,\ldots\ldots, N-1$

$x(n)$ be the mapped sequence and

N be the length of the sequence.

Step-3: Calculate mean μ_h, μ_c, μ_t and variances $\sigma_h, \sigma_c, \sigma_t$ of $X_h(k)$, $X_c(k)$ and $X_t(k)$ respectively.

Step-4: Find out PDF of each sequence using Gaussian Distribution function using equation (2).

$$X(k) = f(x \mid mean, variance) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (2)$$

where $X(k)$ = Spectrum of sequence, x =the sequence, μ= mean of the sequence and σ= variance of the sequence.

Step-5: Apply Bayesian Fusion technique on the mean μ_h, μ_c, μ_t and variances $\sigma_h, \sigma_c, \sigma_t$ using equation (3) and equation (4) to get the fused probability density function.

$$\mu_f = (X_h(k) \times \mu_h + X_c(k) \times \mu_c + X_t(k) \times \mu_t) / (k_h + k_c + k_t) \quad (3)$$

$$\sigma_f = ((X_h(k) \times \sigma_h^2 + X_c(k) \times \sigma_c^2 + X_t(k) \times \sigma_t^2) + (X_h(k) \times d_h^2 + X_c(k) \times d_c^2 + X_t(k) \times d_t^2)) / ((k_h + k_c + k_t)/3) \quad (4)$$

where μ_f = fused mean ,

σ_f = fused variance

d_h = difference between μ_f and μ_h

d_c = difference between μ_f and μ_c

d_c = difference between μ_f and μ_t

k_h, k_c, k_t be the length of healthy, cancer and target genes respectively.

Step-6: Plot spectrum of fused probability density function using equation (5).

$$X_{BF}(k) = f(x_f \mid mean, variance) = \frac{1}{\sigma_f\sqrt{2\pi}} e^{\frac{-(x_f-\mu_f)^2}{2\sigma_f^2}} \quad (5)$$

where $X_{BF}(k)$ = Spectrum of Bayesian fused sequence
X_f= Bayesian fused sequence
μ_f = fused mean and
σ_f= fused variance

3. RESULT AND DISCUSSION

In the present paper the authors have tested the algorithm on several healthy and cancer genes of prostate, breast, colon and gastric cells and MATLAB (2009 b) environment is used to simulate the results. DFT Spectrum of both cancer and healthy gene sequences, shown in Fig. 2 which is random in nature and no decision can be taken from the plots.

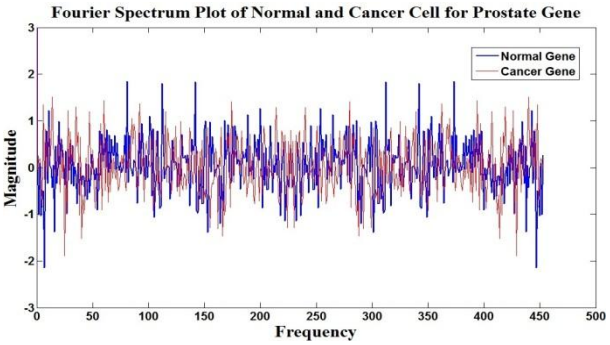

Fig. 2 Comparison between FFT Plots of Normal and Cancer Cells of Prostate Gene

Whenever Bayesian Fusion technique is applied on spectrum of genes, prediction can be taken more accurately from the spectrum. In this method one healthy gene (AF331165.1) and one cancer gene (NP001035756.1) are taken as references in case of Prostate gene. When target gene (healthy) is fused using Bayesian Fusion algorithm, gives a single positive spectrum as shown in Fig. 3a, 4a, 5a, 6a. And while target gene (cancer) is fused using the same, Bayesian Fusion

technique gives a single negative spectrum as shown in Fig. 3b, 4b, 5b, 6b. This spectrum feature can be treated as signature to identify cancerous or healthy genes. Therefore, Bayesian Fusion model proposed by the authors provide more clear identification of genes (cancer or healthy).

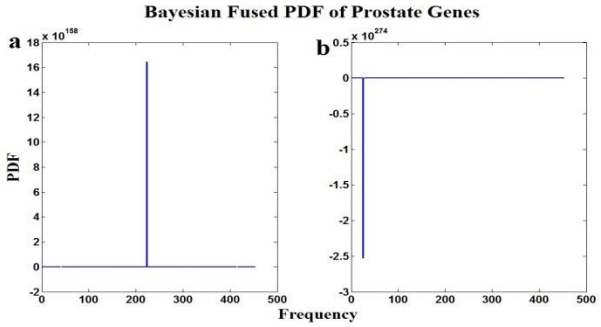

Fig. 3 (a) Plot of Bayesian Fused PDF of Prostate Cancer gene (NP001035756.1), Healthy gene (AF331165.1) and Healthy gene (AF224278.1), (b) Plot of Bayesian Fused PDF of Cancer gene (NP001035756.1), Healthy gene (AF331165.1) and Cancer gene (AAQ08976.1)

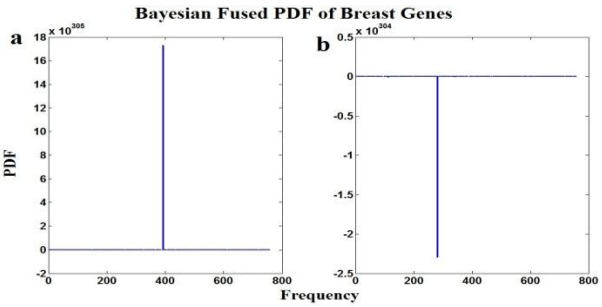

Fig. 4 (a) Plot of Bayesian Fused PDF of Breast Cancer gene (NM_007298.3), Healthy gene (NM_001206932.1) and Healthy gene (NM_001206936), (b) Plot of Bayesian Fused PDF of Cancer gene (NM_007298.3), Healthy gene (NM_001206932.1) and Cancer gene (NM_007299.3)

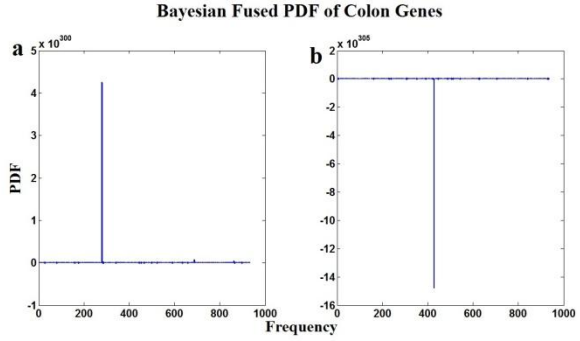

Fig. 5 (a) Plot of Bayesian Fused PDF of Colon Cancer gene (AY572973.1), Healthy gene (NM_017436.4) and Healthy gene (NM_001173466.1), (b) Plot of Bayesian Fused PDF of Cancer gene (AY572973.1), Healthy gene (NM_017436.4) and Cancer gene (AY217549.1)

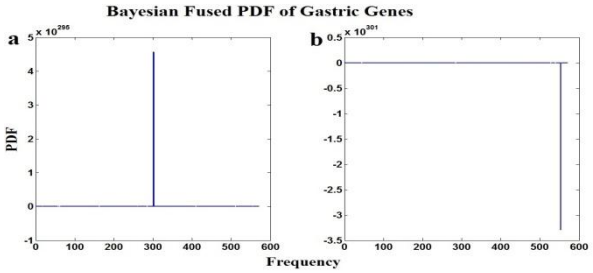

Fig. 6 (a) Plot of Bayesian Fused PDF of Gastric Cancer gene (NM_000108.3), Healthy gene (NM_001909.4) and Healthy gene (NM_012245), (b) Plot of Bayesian Fused PDF of Cancer gene (NP001035756.1), Healthy gene (NM_001909.4) and Cancer gene (NM_0006135.2)

The peak value of the amplitude spectrum for fused cancer and healthy gene are listed in Table 2, 3, 4, 5.

Table 2. List of Peak Amplitudes of Fused Spectra with Their Corresponding Accession no.

Sl. No.	Accession No.	Amplitude of the Spectrum
Prostate Normal Cell		
1	AF224278.1	1.644e+159
2	NM007003.2	2.228e+163
3	NM_005984.3	3.694e+221
4	M24902.1	1.739e+254
5	M24543.1	2.370e+126
6	M15885.1	2.371e+307
Prostate Cancer Cell		
1	AAQ08976.1	-2.527e+274
2	NP001231873.1	-1.035e+295
3	FJ649644.1	-5.492e+216
4	AY008445.1	-2.626e+300
5	AF455138.1	-5.535e+297
6	AF304370.1	-1.698e+194

Table 3. List of Peak Amplitudes of Fused Spectra with Their Corresponding Accession no.

Sl. No.	Accession No.	Amplitude of the Spectrum
Breast Normal Cell		
1	NM_001206932.1	1.556e+299
2	NM_001206934	5.21e+307
3	NM_001206936.1	2.451e+307
4	NM_001206940.1	1.807e+302
5	NM_001206954	2.703e+302
6	NM_001206966	2.986+293
Breast Cancer Cell		
1	NM_007298	-1.215e+305
2	NM_014567.3	-5.071e+282
3	NM_001170717	-4.07e+304
4	NM_001170718	-1.201e+307
5	NM_00170721	-1.191e+307
6	NM_001261408.1	-4.32e+302

Table 4.List of Peak Amplitudes of Fused Spectra with Their Corresponding Accession no.

Sl. No.	Accession No.	Amplitude of the Spectrum
Colon Normal Cell		
1	NM_015665	2.076e+304
2	NM_001173466.1	4.255e+300
3	NM_207365.3	1.763e+302
4	NM_207365.3	7.074e+302
5	NM_153698.1	3.293e+302
6	NM_001087.3	3.222e+303
Colon Cancer Cell		
1	AB 489153.1	-1.48e+306
2	AY 572973.1	-6.098e+303
3	AY 601851.1	-1.692e+308
4	AY 217549.1	-2.535e+307
5	AF 256731.1	-3.012e+297
6	AY 581148.1	-2.815e+303

Table 5. List of Peak Amplitudes of Fused Spectra with Their Corresponding Accession no.

Sl. No.	Accession No.	Amplitude of the Spectrum
Gastric Normal Cell		
1	NM_12245	4.57e+295
2	NM_004181.4	1.48e+224
3	XM_005267414	1.339e+293
Gastric Cancer Cell		
1	NM_0006135.2	-3.29e+301
2	NM_000291.3	-6.514e+286
3	NM_013283.4	-1.534e+307

4. COMPARISON WITH EXISTING METHOD

The authors S. Barman (Mandal) and et al [13], in their paper, estimated power spectrum of genes and filtered for identification of healthy and cancer genes. They calculated ratio of mean amplitude to mean normalized frequency and claimed the ratio is more than 1 for cancer cell whereas for healthy cells less than 1. But it has been observed that the ratio does not work truly for all types of genes. Whenever we applied this method on prostate, breast, colon and gastric genes, it shows a difference in ration for healthy and cancer genes only in case of colon and breast genes (Fig.8 and 9). In case of other genes like prostate and gastric no conclusion can be made based on the ratio of mean amplitude to mean normalized frequency (Fig. 7 and 10). Therefore, a robust method is a challenge to the researchers which can screen out cancer genes from healthy genes. The proposed Bayesian Fusion Technique shows a marked difference in Probability Density Function for cancer and healthy genes not only for breast and colon but also for prostate and gastric genes.

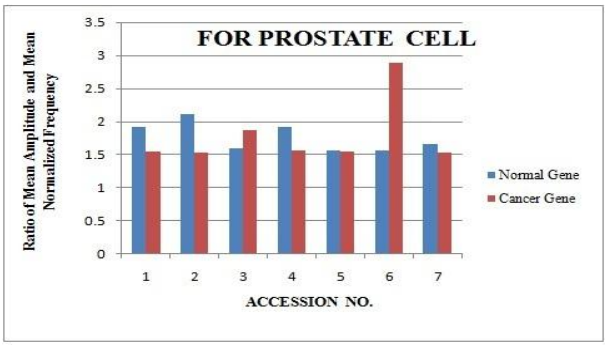

Fig.7 Comparison of Ratio of Mean Amplitude and Mean Normalized Frequency between Healthy and Cancer Cell of Prostate Genes

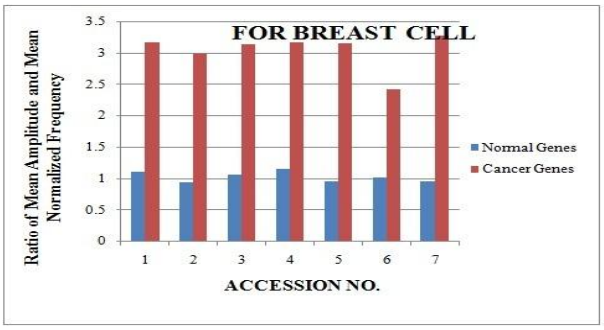

Fig.8 Comparison of Ratio of Mean Amplitude and Mean Normalized Frequency between Healthy and Cancer Cell of Breast Genes

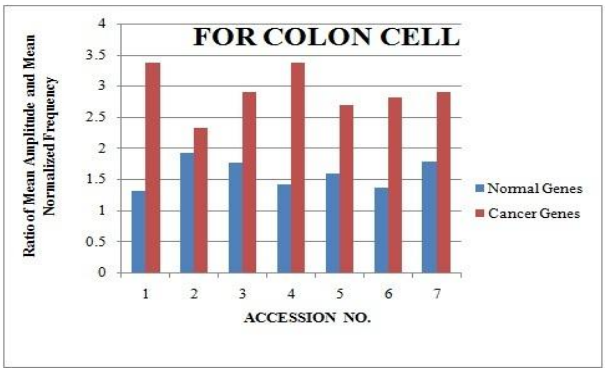

Fig.9 Comparison of Ratio of Mean Amplitude and Mean Normalized Frequency between Healthy and Cancer Cell of Colon Genes

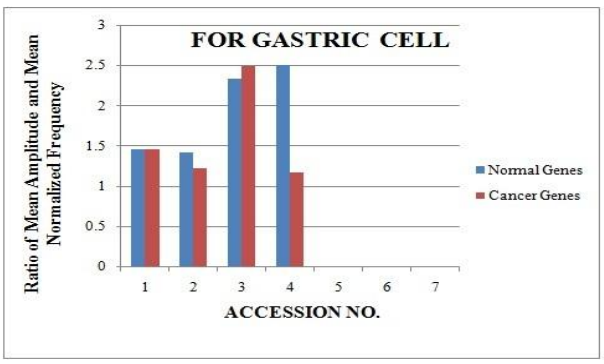

Fig.10 Comparison of Ratio of Mean Amplitude and Mean Normalized Frequency between Healthy and Cancer Cell of Gastric Genes

The ratio of mean amplitude to mean normalized frequency of normal and cancer genes of prostate, breast, colon and gastric cells are listed in Table. 6, 7, 8 and 9. The authors showed here positive and negative spectrum based decision rule which is more appropriate in case of classifier design issue.

Table 6. List of Ratio of Mean Amplitude and Mean Normalized Frequency with their corresponding Accession No. for Prostate Genes

Sl. No.	Accession No.	Ratio of Mean Amplitude and Mean Normalized Frequency
Prostate Normal Genes		
1	NM007003.2	1.9127
2	NM_005984.3	2.1041
3	M24902.1	1.5829
4	M24543.1	1.9082
5	M15885.1	1.5534
6	AF462605.1	1.5556
7	AF331165.1	1.6598
Prostate Cancer Genes		
1	NP001231873.1	1.5446
2	NP001035756.1	1.5201
3	FJ649644.1	1.8715
4	AY008445.1	1.5563
5	AF455138.1	1.5322
6	AF304370.1	2.8844
7	AAQ08976.1	1.5185

Table 7. List of Ratio of Mean Amplitude and Mean Normalized Frequency with their corresponding Accession No. for Breast Genes

Sl. No.	Accession No.	Ratio of Mean Amplitude and Mean Normalized Frequency
Breast Normal Genes		
1	NM_001206929.1	1.1060
2	NM_001206932.1	0.9332
3	NM_001206934.1	1.0475
4	NM_001206936.1	1.1482
5	NM_001206940.1	0.9456
6	NM_001206954.1	1.0092
7	NM_001206966.1	0.9456
Breast Cancer Genes		
1	NM_007298.3	3.1600
2	NM_007299.3	2.9829
3	NM_014567.3	3.1387
4	NM_001170717	3.1674
5	NM_001170718	3.1537
6	NM_001170721	2.4110
7	NM_001261408.1	3.2769

Table 8. List of Ratio of Mean Amplitude and Mean Normalized Frequency with their corresponding Accession No. for Colon Genes

Sl. No.	Accession No.	Ratio of Mean Amplitude and Mean Normalized Frequency
Colon Normal Genes		
1	NM_017436.4	1.3062
2	NM_015665.5	1.9230
3	NM_001173466.1	1.7711
4	NM_001086.2	1.4138
5	NM_207365.3	1.5906
6	NM_153698.1	1.3563
7	NM_001087.3	1.7740
Colon Cancer Genes		
1	AB489153.1	3.3653
2	NM_001167617.1	2.3237
3	AY572973.1	2.8934
4	AY601851.1	3.3635
5	AY217549.1	2.6885
6	AF250731.1	2.8129
7	AY572972.1	2.8933

Table 9. List of Ratio of Mean Amplitude and Mean Normalized Frequency with their corresponding Accession No. for Gastric Genes

Sl. No.	Accession No.	Ratio of Mean Amplitude and Mean Normalized Frequency
Gastric Normal Genes		
1	NM_001909.4	1.4574
2	NM_004181.4	1.4145
3	NM_012245	2.3336
4	XM_005267414.1	2.4960
Gastric Cancer Genes		
1	NM_000108.3	1.4525
2	NM_000291.3	1.2163
3	NM_006135.2	2.4864
4	NM_013283.4	1.1667

5. CONCLUSION

A combination of Discrete Fourier Transform and Bayesian Fusion change weak prediction to reliable prediction of cancer and healthy genes. The method is successfully tested on genes collected from NCBI homepage which can be used to separate out cancer genes or healthy genes from enormous gene database. The implementation of the approach is simple as binary decision rule (two logic based) is used for detection of healthy or cancer gene. Our results showed it works truly for any genes like breast, prostate, colon and gastric. In order to find robust detection approach, it is necessary to explore the best combination of different fusion strategies and search the intrinsic association between differentfusion techniques and compare their performance in the future.

6. REFERENCES

[1] D.Anastassiou., "Genomic Signal Processing." IEEE Signal Processing Magazine (2001); pp.8-20.

[2] P.P Vaidyanathan,.and B.J. Yoon, "The role of signal-processing concepts in genomics and proteomics," Journal of the Franklin Institute, 341.1(2004); pp.111-135.

[3] S.Saha and S. Barman(Mandal), "Digital filtering of Amino acid sequence for prediction of cancer cell", 2nd Annual International Conference on Electronics Engg. & Computer Science(IEMCON 2012).

[4] A. Ghosh, and S. Barman."Prediction of Prostate Cancer Cells based on Principal Component Analysis Technique."Procedia Technology vol.10 (2013); pp. 37-44.

[5] S.Barman(Mandal), S.Saha, A.Mandaland M. Roy, "Signal Processing Techniques for the analysis of Human Genome associated with cancer cells", International Conference on Electronics Engg. & Computer Science(IEMCON2011) Organized in collaboration with IEEE, INDIA, pp. 570-573

[6] M.Roy and S. Barman (Mandal),"Application of Principal Component-Minimum Variance Technique in Gene Prediction", Review of Applied Physics Vol.2 Iss.4(2013); pp.106-113

[7] X. Dai,O.Yli-Harja and H.Lahdesmaki, "Novel Data Fusion Method and Exploration of Multiple Information Sources for Transcription Factor Target Gene Prediction", EURASIP Journal on Advances in Signal Processing, Volume 2010, Article ID 235795

[8] A.Ross and A. Jain, "Information fusion in biometrics", Pattern Recognition Letters 24 (2003) 2115–2125, Elsevier Science

[9] H. Liu, D.Yue, L. Zhang, Y. Chen, S.J.Gao and Y. Huang, "A Bayesian approach for identifying miRNA targets by combining sequence prediction and gene expression profiling", International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing (IJCBS), Shanghai, China. 3-8 August 2009

[10] I. M. EI-Badawy, A. M. Aziz, S. Gasser and M. E. Khedr, "A New Multiple Classifiers Soft Decisions Fusion Approach for Exons Prediction in DNA Sequences", IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2013

[11] M. Raza, I. Gondal, D. Green, and R. L. Coppel, "Fusion of FNA-cytology and Gene-expression Data Using Dempster-Shafer Theory of Evidence to Predict Breast Cancer Tumors", Bioinformation 1.5 (2006): 170

[12] P. Ray, L. Zheng, J. Lucas and L. Carin, "Bayesian joint analysis of heterogeneous genomics data", Bioinformatics, Vol. 30 no. 10 2014, pages 1370–1376

[13] S.Barman (Mandal), M. Roy, S.Biswas and S. Saha, "Prediction of Cancer Cell using Digital Signal Processing", Annals of Faculty Engineering Hunedoara (International Journal of Engineering) , ISSN 1584-2673 (2001).

[14] E. R. Dougherty, A. Datta, and C. Sima, "Research Issues in Genomic Signal Processing", IEEE Signal Processing Magazine, November (2005):pp.46-68

[15] T.Seok Jin, J. Myung Lee and S. K. Tso, "A new approach using sensor data fusion for mobile robot navigation," Int. Journal Robotica vol. 22(2004); pp. 51–59

[16] M.Kumar, D.P. Garg, and Randy A. Zachery, "A Method for Judicious Fusion of Inconsistent Multiple Sensor Data," in IEEE Sensors Journal,Vol.7, no.5,(2007)

[17] J.Z.Sasiadek, "Sensor Fusion", Annual Reviews in Control 26 (2002), Elsevier Science,pp.203-228

[18] Thrun, Burgard,"Probabilistic Robotics", Fox, MIT Press, Cambridge.

[19] National Centre for Biotechnology Information (NCBI). (http://www.ncbi.nlm.nih)

[20] S. Achuthsankar Nair and S. Pillai Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)", ISSN 0973-2063, Bioinformation 1(6): pp.197-202