Original Article

ISSN: 2349 - 4891

International
Journal of Recent Research and Applied Studies
(Multidisciplinary Open Access Refereed e-Journal)

Comparison of Correlation Method with Modified STFT Method for Breast Cancer Specific mRNA-TF Interaction analysis

Binthiya Suny Gabriel[1]&DrTessamma Thomas[2]

[1]*Research Scholar, Department of Electronics, Cochin University of Science and Technology, Kochi, Kerala, India 682022*
[2]*Research Supervisor, Retd. Professor,Department of Electronics, Cochin University of Science and Technology, Kochi, Kerala, India 682022*

Received 02ndFebruary2021, Accepted 9th August 2021

Abstract

Breast cancer is a sophisticated disease and a detailed research into the mechanisms underlying the development of tumour, has paved a way towards the characterization of Transcription Factors (TFs). The application of TFs in the breast cancer therapeutics is an area of interest to several researchers. A way to bring out the regulatory relations between TF and its target genes, messenger RNA (mRNA), is to measure the changes in the expression of the gene in response to TF perturbation. TFs are considered to regulate the expression of more than 20 per cent of the entire gene in the mammalian cells. The mRNAs with the greatest change in the expression levels are not necessarily the ones that are most relevant. However, differentially expressed TFs have greater importance in a biological context in relation to the progression of cancer than TFs that target and modulate just a few mRNA transcripts. In this paper, a new technique for analysing this nature of modulation by TF, has been developed, which determines the binding regions of TFs to the mRNAs using a normalised correlation method. The analysis includes 30 breast cancer specific mRNAs and the various TFs that target each mRNAs. The new correlation method identifies the binding regions of all the TFs and the results obtained are comparable with those obtained for the STFT method.

KEYWORDS; mRNA, Breast Cancer, TF, Binding Region, Normalised Correlation.

Introduction

Transcription Factors (TFs) are proteins involved in the process of converting DNA into RNA. One special feature of TFs is that they have DNA binding domains that give them the advantage of binding to certain sequences of DNA known as the promoter or the enhancer sequences. Some transcription factors bind to the promoter sequence of the DNA near the site where the transcription starts and help form the transcription initiation complex. Other transcription factors bind to the regulatory sequences, such as enhancer sequences, and can either repress or stimulate transcription of the related gene [1]. Transcription Factors (TFs) play a major role in the gene expression regulation that yields to different levels of proteins and gene transcripts. Due to the non-availability of direct technological platforms to analyse these complex interactions of TF with both miRNA and mRNA, the integration of different datasets has gained much importance. Consideration of TFs brings a whole new aspect to the miRNA-gene networks as TFs can regulate both genes and miRNAs thereby increasing the number of gene to miRNA and gene to gene interactions. Several strategies are available for analysing the TF-Gene interaction. One such strategy is to consider the interactions present in available and known databases whereas the other techniques depend on the co-expression analysis present in interaction prediction based on the profiles of the expressions present [2]. Both the methods mentioned have limitations. Analysis done at the database level lacks tissue-specificity. Moreover, the number of interactions declines when limited to one cell type. The analysis done on the basis of correlative studies can result in a high false positives rate. This is because a high value in the correlation of the expression does not offer a guaranteed interaction between TF and the target gene. In this paper, the new correlation method is applied for obtaining the binding regions of several TFs for 30 breast cancer specific mRNAs. The results obtained for this method is same as the results obtained from the already existing method for analysing the TF-target gene interaction; the STFT method.

BIOLOGICAL BACKGROUND

Transcription factors

Transcription Factors are proteins that helps to turn specific genes "on" or "off" by binding to nearby DNA.

Correspondence Author
Binthiya Suny Gabriel, *Research Scholar of Department of Electronics, CUSAT.*

Activators are transcription factors that boost a gene's transcription. **Repressors** decrease the level of transcription. Groups of transcription factor binding sites called silencers and **enhancers** can turn a gene off/on in specific body parts. Transcription factors allow cells to perform logic operations and combine different sources of information to "decide" whether to express or supress a gene. Transcription Factors help ensure that the right genes are expressed in the right cells of the body, at the right time.

Working mechanism of Transcription Factors

A typical transcription factor binds to a specific target sequence within the DNA after which the TF makes it either harder or easier for RNA polymerase to bind to the promoter of the gene.

1. Activators

Activators are transcription factors that may help the general transcription factors or RNA polymerase to bind to the promoter region. Figure 1 [3] shows how the general transcription factors get help from the activators in order to assemble.

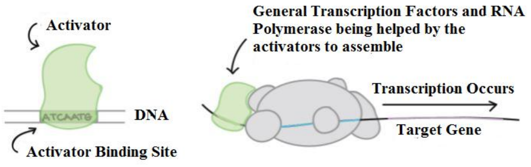
Activator
General Transcription Factors and RNA Polymerase being helped by the activators to assemble
Transcription Occurs
DNA
Target Gene
Activator Binding Site

Figure 1. RNA polymerase and General Transcription Factors binding to the promoter region with the help of activators

2. Repressors

Few transcription factors repress transcription. One simple way by which transcription gets repressed is by getting in the way of the general transcription factors or the RNA polymerase in such a way that the binding to the promoter becomes hard or the initiation of the transcription becomes difficult. Figure 2 [3] shows how the repressors alter or hinder the transcription process.

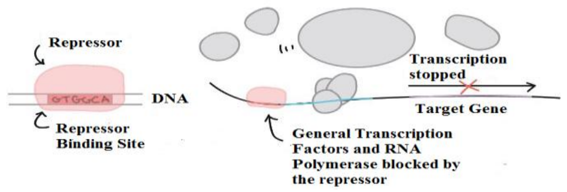
Repressor
DNA
Repressor Binding Site
Transcription stopped
Target Gene
General Transcription Factors and RNA Polymerase blocked by the repressor

Figure 2. RNA polymerase and General Transcription Factors blocked by the repressor

3. Binding Sites

The transcription Factor binding sites are close to the promoter of a gene. However, these sites could also be found at a distance, very far away from the promoter, in which case also the transcription can be affected. The activation domain binds to the mediator and helps in the start of the transcription process.

METHOD

Recent findings have elucidated that the regulation of messenger RNA (mRNA) levels is due to the synergistic and antagonist actions of transcription factors (TFs) and microRNAs (miRNAs). Mutual interactions among these molecules are easily modelled and analysed using several techniques.

1. Graphs and Sub Graphs: The application of graphs and sub graphs like feed forward loops or regulatory loops is one among the many techniques used for analysing the interaction. The technological platforms that are currently available aid in the analysis of only one aspect of these mechanisms [4].

However the analysis and integration of various data sources makes it possible to perform a comprehensive analysis. A lot of information has been made available through the classical approaches of analysis about the application of single class molecules. However there is a lack of novel techniques being introduced in order to analyze the interplay of molecules by the integration of these data sources into a single comprehensive one [5]. *This integration only aids in explaining how these networks regulate the diseases and biological processes at the systems level.*

2. Genomic Signal Processing techniques: The TF-mRNA interaction is greatly influenced by the binding strength between the two at the region of the seed. The role of the seed sequence is significant in the analysis of the extent of interaction between TF and mRNA towards cancer. The importance of signal processing methods is attributed to their application in processing and interpreting the information present in genomics data. Several DSP based algorithms have been applied to studying the characteristics of DNA and RNA sequences. The importance of genomic signal processing techniques is rising as it has been recognized that the characterization of genomic and proteomic regulations require various disciplines related to signal processing [6]. Several Digital Signal Processing techniques including DFT and STFT have found application in the search for genomic repeats using Fourier analysis. DFT was used for spectrum analysis of biological data where initially the DNA sequence was mapped into a numeric sequence and spectrum of finite-length windowed DNA numerical sequences was computed. The applications of digital filters also helped partially eliminate the background 1/f noise of the spectrum exhibited by nearly all DNA sequences [7]. *However, the solutions provided by these algorithms include much background noise and the results obtained are computationally complex and less accurate. Additional techniques are needed to remove the noise.* When some of the TFs bind with mRNAs, degradation may occur, leading to cancer.

3. The new proposed method: Our purpose is to find out such binding regions and seed binding regions specific to breast cancer, without the presence of any background noise. *Normalised Correlation method is used for this purpose.*

MATERIALS

Database

The program codes for the normalised correlation method, was designed using the built-in MATLAB utility. The lists of the TFs that target a specific mRNA are taken from the NCBI website and the sequences of the TFs are obtained from the TRANSFAC website [https://academic.oup.com/nar/article/24/1/238/2360291] The mRNA sequences are obtained from the UCSC Genome browser website [https://www.genome.ucsc.edu/]. The breast cancer specific mRNAs EGFR, and BRCA1, and the TFs that target these three mRNAs were used for analysis, in this study. **3**

EXPERIMENT

When some of the TFs bind with mRNAs, degradation may occur, resulting in cancer. Our purpose is to detect binding regions specific to breast cancer, without any noise in the background. Normalised Correlation method has been used for this purpose.

1. Correlation Method for determining TF binding regions

The different TFs and the breast cancer specific mRNAs whose regions of binding to the TFs are to be found, are selected. The calculation of maximum correlation between TF and the mRNA initially needs the reversed compliment of the TF sequence to be found [8]. The normalised correlation between mRNA and TF was computed while laterally shifting TF throughout the mRNA sequence and calculating the normalised correlation value each time the TF sequence is shifted. The maximum value of the correlation strength from among the set of correlation values obtained, is noted. The only requirement would be to do a one-to-one match for a short length of bases between the reverse complemented TF and the mRNA in order to get the binding region. As per the normalised correlation method, the binding regions of various TFs are obtained for 30 breast cancer specific mRNAs. In this paper, the results obtained for the breast cancer specific mRNAs, EGFR, and BRCA1 are tabulated in Tables I, and II respectively.

EGFR: EGFR is a critical factor which actively participates in the occurrence and progress of Non-Small Cell Lung Cancer (NSCLC). EGFR is seen to be overexpressed in a lot of patients having NSCLC. This is considered as an important target in the treatment for cancer [9].

BRCA1: The mutations in the BRCA1 gene are found to be a major cause of breast cancer and thus supresses tumour. Moreover, this gene plays a major role in the stability of the genome. With mutations in BRCA1, the risk of developing breast cancer is 80%. Investigations done on BRCA1 for various organisms have provided knowledge of the involvement of BRCA1 in breast cancer [10].

2.Implementation of the Correlation Method

According to the normalised correlation method, the TF binding regions are obtained for 30 breast cancer specific mRNAs of which the TF binding regions of EGFR, and BRCA1 are presented in this paper and tabulated in tables I, II respectively.

3. Modified STFT Method

Maggi et al. [11] considered applying the STFT to the indicator sequence. The indicator sequence is obtained by replacing a nucleotide with its corresponding EIIP (Electron Ion Interaction Potential) value and selecting every sixth component to obtain the spectral content. In the modified STFT method, a normalizing technique is applied to the spectral content by dividing with the total spectral value. This technique is then followed by scaling, to obtain the peaks which give the binding regions. The regions where the binding occurs (binding regions), is selected by noting the value of only those peaks and the position of the peaks, which lie above the standard deviation of the total spectral content value.

RESULTS AND DISCUSSION

The number of binding regions obtained for EGFR, and BRCA1 are 20, 24 respectively, for the study. The comparative details of the binding regions with respect to the modified STFT and correlation methods, for EGFR, and BRCA1 respectively, have been shown in Tables 1, and 2. Considering the correlation method, the P53 TF sequence (of length 86) binding to EGFR (Table 1), have nearly the same binding regions as that obtained using the Modified STFT method.

Table 1. Modified STFT and Correlation Method results obtained for the interaction between EGFR mRNA and TP53 Transcription Factor

EGFR	***Modified STFT Method***		***Correlation Method***	
Sl. No.	Peak Position, Peak Strength at the peak position	Peak Width	Base Position at which the Maximum Correlation Strength is obtained, Maximum Correlation Strength	Binding Region
1	8195, 0.1093	8153-8238	8195, 0.848079	8153-8238
2	100, 0.114	58-143	100, 0.9568	58-143
3	6266, 0.1415	6224-6309	6265, 0.9595	6223-6308
4	651, 0.1438	609-694	650, 0.9608	608-693
5	4729, 0.1707	4686-4772	4728, 0.9611	4686-4771
6	653, 0.2186	611-696	653, 0.9619	611-696
7	4369, 0.2343	4327-4412	4369, 0.9624	4327-4412
8	7332, 0.254	7290-7375	7332, 0.965	7290-7375
9	7990, 0.4066	7946-8032	7989, 0.966	7947-8032
10	27, 0.427	13-99	27, 0.9675	13-99
11	5027, 0.5111	5011-5097	5026, 0.9688	5011-5096
12	966, 0.5558	936-1021	966, 0.9695	936-1021
13	9390, 0.5801	9111-9196	9390, 0.9701	9111-9196
14	2254, 0.9466	2235-2321	2253, 0.9713	2237-2321
15	6776, 1.055	6720-6806	6775, 0.9723	6721-6806
16	2645, 1.2	2625-2711	2644, 0.9734	2626-2711
17	2014, 1.232	1985-2071	2014, 0.9748	1986-2071
18	8889, 1.244	8827-8913	8889, 0.9765	8828-8913
19	8619, 1.328	8613-8699	8618, 0.9795	8614-8699
20	3652, 2.91	3544-3630	3652, 0.9803	3545-3630

In the results obtained using the correlation method, the 20 TFs binding to EGFR (Table 1), have exactly the same binding position as that in the case of Modified STFT method. However, few of the binding regions have length less by 1 base.

ISSN: 2349 – 4891

Table 2. Modified STFT and Correlation Method results obtained for the interaction between BRCA1 mRNA and TP53 Transcription Factor

BRCA1	***Modified STFT Method***		***Correlation Method***	
Sl. No.	Peak Position, Peak Strength at the peak position	Peak Width	Base Position at which the Maximum Correlation Strength is obtained, Maximum Correlation Strength	Binding Region
1	334, 0.6362	291-377	333, 0.9086	290-376
2	753, 0.2461	710-796	753, 0.9099	710-796
3	1081, 0.7086	1038-1124	1082, 0.9195	1039-1125
4	1220, 0.3099	1177-1263	1220, 0.9198	1177-1263
5	1682, 0.4908	1639-1725	1683, 0.9232	1640-1726
6	2004, 1.198	1961-2047	2006, 0.9237	1963-2049
7	2098, 0.2503	2055-2141	2098, 0.925	2055-2141
8	2374, 1.668	2331-2417	2374, 0.9278	2331-2417
9	2455, 0.3616	2412-2498	2456, 0.9308	2413-2499
10	2608, 0.7077	2565-2651	2609, 0.9361	2566-2652
11	2990, 0.2324	2947-3033	2990, 0.9362	2947-3033
12	3248, 0.2784	3205-3291	3249, 0.9435	3206-3292
13	3356, 0.2373	3313-3399	3356, 0.9461	3313-3399
14	3417, 0.5443	3374-3460	3418, 0.9473	3375-3461
15	3856, 0.5413	3813-3899	3856, 0.9488	3813-3899
16	3958, 1.308	3915-4001	3957, 0.949	3914-4000
17	4200, 0.6734	4157-4243	4200, 0.9544	4157-4243
18	4448, 0.6804	4405-4491	4449, 0.9544	4406-4492
19	4731, 0.4357	4688-4774	4731, 0.9645	4688-4774
20	5425, 0.3842	5382-5468	5424, 0.9668	5381-5467
21	5705, 0.473	5662-5748	5704, 0.9673	5661-5747
22	5882, 2.488	5839-5925	5882, 0.968	5839-5925
23	6156, 0.4645	6113-6199	6157, 0.9684	6114-6200
24	6293, 1.195	6250-6336	6293, 0.9707	6250-6336

Similarly, in Table 2, considering the correlation method, the P53 TF sequence (of length 86) binding to BRCA1 (Table 2), have nearly the same binding regions as that obtained using the Modified STFT method except in few instances where the binding region varies by one nucleotide base.

Computational load

The computation in modified STFT method required is more; FFT calculation is essential and is applied to the indicator sequence using MATLAB. In this method, the binding region obtained is approximate. However, in correlation method, the binding regions are obtained at the point of maximum correlation.

CONCLUSIONS

Out of the methods that were analysed namely, modified STFT method, and correlation method, for detecting the miRNA binding regions for the breast cancer specific mRNAs, EGFR and BRCA1, correlation method provided results exactly the same as that in ground truth. The modified STFT method, gives peaks, corresponding to all the approximate binding regions.

FUTURE SCOPE

The two methods considered in this study, have provided results that could help in further analysing the nature of the TF and mRNA interaction. Identifying the seed region is also crucial in evaluating this interaction. The correlation method has advantage over the modified STFT method in finding not only the binding regions but also the seed binding regions of TF. Identifying the seed region is crucial in evaluating the extent of breast cancer progression [12]. The next phase of this research would include finding the relation between seed length and the complexity of breast cancer cells.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support from Department of Electronics, Cochin University of Science and Technology, Kochi, Kerala.

REFERENCES

[1] Latchman D.S. (1993). Transcription factors: an overview. *International Journal of experimental pathology*, *74*(5), 417–422.

[2] Nersisyan S, Galatenko A, Galatenko V, Shkurnikov M, Tonevitsky A (2021) miRGTF-net: Integrative miRNA-gene-TF network analysis reveals key drivers of breast cancer recurrence. PLOS ONE 16(4): e0249424.

[2]Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, G.P., & Eliopoulos, E. (2020). Transcription factors and evolution: An integral part of gene expression (Review). World Academy of Sciences Journal, 2, 3-8. https://doi.org/10.3892/wasj.2020.32

[3]Transcriptionm factors. https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/regulation-of-gene-expression-and-cell-specialization/a/eukaryotic-transcription-factors.

[4] Guzzi, P. H., Di Martino, M. T., Tagliaferri, P., Tassone, P., &Cannataro, M. (2015). Analysis of miRNA, mRNA, and TF interactions through network-based methods. EURASIP journal on bioinformatics & systems biology, *2015*, 4.https://doi.org/10.1186/s13637-015-0023-8.

[5] Ernst, J., Plasterer, H. L., Simon, I., & Bar-Joseph, Z. (2010). Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome research*, *20*(4), 526–536.https://doi.org/10.1101/gr.096305.109.

[6] Bayat, A. (2002). Science, medicine, and the future: bioinformatics. US National Library of Medicine: National Institute of Health, 324(7344), 1018-1022.

[7] Serpedin, E., Garcia-Frias, J., Huang, Y., and Braga-Neto, U. (2009). Applications of signal processing techniques to bioinformatics, genomics, and proteomics. EURASIP Journal on Bioinformatics and Systems Biology, Article number: 250306, 1-2.

[8] Mullany, L.E., Herrick, J.S., Wolff, R.K., and Slattery, M.L. (2016). MicroRNA seed region length impact on target messenger RNA expression and survival in colorectal cancer. PLOS ONE, 11(4): e0154177.

[9] Tasdemir, S., Taheri, S., Akalin, H., Kontas, O., Onal, O., &Ozkul, Y. (2019). Increased EGFR mRNA Expression Levels in Non-Small Cell Lung Cancer. The Eurasian journal of medicine, 51(2), 177–185.https://doi.org/10.5152/eurasianjmed.2016.0237.

[10] Gudmundsdottir, K., Ashworth, A. (2006). The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene* **25,** 5864–5874.

[11] Maggi, N.; Arrigo, P.; and Ruggiero, C. (2013). Optimize ncRNA targeting: a signal analysis based approach. IFMBE Proceedings of the XIII Mediterranean Conference on Medical and Biological Engineering and Computing. Cyprus, 662-665.

[12] Huang, J.C.; Babak, T.: Corson, T.W.; Chua, G.; Khan, S.; Gallie, B.L.; Hughes, T.R.; Blencowe, B.J.; Frey, B.J.; and Morris, Q.D. (2007). Using expression profiling data to identify human microRNA targets. Nature Methods, 4, 1045-1049.

.