ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Special Issue (April)

Online - 2455-3891
Print - 0974-2441

Full Proceeding Paper

PREDICTION OF CHRONIC BACTERIAL INFECTION BY IDENTIFICATION OF INTERCELLULAR RESPONSES OF GENETIC FUSION CENTERS

ANKUSH RAI, JAGADEESH KANNAN R

School of Computing Science and Engineering, Chennai, Tamil Nadu, India. Email: ankushressci@gmail.com

Received: 28 December 2016, Revised and Accepted: 10 May 2017

ABSTRACT

Objective: In this study, we have designed an algorithm for early detection of DNA fusion to discover the potential transcription which embodies the fusion of gene products derivable from the human DNA with that of bacterial and cancerous viruses, resulting from the several breakage points and re-assembling of different chromosomes, or that of within a chromosome.

Methods: Without relying on existing annotations, the proposed algorithm proves its efficacy in detecting alignment of RNA sequences from unannotated splice variants of known genome strands. Using this algorithm in the age of Big Data analytics the potential threat of cancer, tuberculosis, tumors, and asthma can be predicted beforehand while scaling such effects, ranging from individual to population scale.

Results and Conclusion: We have also reported the results of the algorithm for over 90 samples with solid supporting evidences and open a new virotherapy approach of a numerically quantized cure for a disease such as cancer, tumors, and asthma.

Keywords: Algorithm, Computational modeling, Gene fusion, DNA transcription.

DOI: http://dx.doi.org/10.22159/ajpcr.2017.v10s1.19978

INTRODUCTION

The choice for the rapid method of quantifying and screening of genes, the RNA-sequencing protocol is highly employed [1-4]. It offers an advantage that it majorly doses not pre-existing knowledge base for the genetic content unlike the sequencing techniques of microarray expression. Thereby, the detection of entirely novel splice variants of existing genes or that of even a completely novel gene fairly easy. But for detection of an entirely novel gene, the software employed to analyze RNA sequence must does not rely heavily on existing annotations and consequentially be capable of aligning the sequence of transcription anywhere in the genome.

In the past few studies, the algorithms by the likes of TopHat and Cufflinks are capable of ab-initio spliced alignment and previously been highly used in sequencing experiments [5,6]. Although it also has the potential of detecting of genes formed by complex of arrangements with chromosomes. However, the detecting of the matching patterns for the fusion genes remains daunting as this fusion gene formed by breakage and linkage of sets of chromosomal arrangements are the primary cause of the arise of cancers, chronic myeloid leukemia, etc., providing a host to bacteria and cancerous virus to rejoin with the vulnerable cells RNA transcription and adversely affects the cell biology [7-9].

There are over 60,000 documented cases have been reported for cancer causes from gene fusion in Mitelman database [10]. However, detecting such fusion will give an advantage of regulation and treatment of dreadful disease but it has remained an overly due task as due to the fact that in a virus affected cells the transcription cells remained the same as that of host and only fraction of it is available and responsible for alteration of transcription as in case of tumors. In addition, due to the availability of huge amount of alternative splice variants produced in a fusion event the non-transcribed promoter element will not be able to detect it. Other methods which are highly dependent on one another for preprocessing and post-processing tasks such as Bowtie, ELAND, SplitSeek, Trans-ABySS, and BLAT depends on known annotation and for searching possible fusion boundaries uses known exons across fusion points [11-20]. To detect such fusion events, we propose the novel special-purpose algorithms from next-generation sequencers. We demonstrate its effectiveness on six different viral and bacterial cell lines, which are populated with multiple gene fusion events, including both known and intermediary novel fusions in due process. This is a very time-consuming process and requires more than 467 CPU hours to compute assembling for 230 threads.

METHODOLOGY: TYPE MATRIX FUSION ALGORITHM (TMFA)

Given an input signal i = (i(1), i(2), i(3),... i(n)) and the true estimation of the true signal $y=(\hat{y}(1)),\hat{y}(2),\hat{y}(3),...)$ is given by:

$$\hat{y}(p)=\frac{1}{N}\sum_{i=0}^{N-1} i(p-1)$$

Thus, for every p≥N, where N is the filter's buffer window for the number of input observation to be fused. Note that N is also the number of steps taken to detect the breaking point in the sequence of two genomes. The input dynamics of the genome sequences is given by the following function:

```
update ( ){
$y(p+1) = \varphi(p)y(p)+G(p) u(p)+w(p) i(p)$
$= H(p)x+v(p)$
}
```

Where φ(p) is the matrix representing the state transition, G(p) is the matrix for input transition, u(p) is the input vector, H(p) is the measurement matrix, and u and v are the backtrack index variables for the intervals of the genomic sequences.

Algorithm: TMFA for early detection of genetic fusion.

0 Evaluate:

$$\hat{y}(p)=\frac{1}{N}\sum_{i=0}^{N-1} i(p-1)$$

```
update( )
Initiate:
    $m(A_1|A_2) = \frac{\sum_{A_1 \cap A_2}^{p} m(\prod_{j=1}^{N} A_j)}{\sum_{A_1 \cap A_2}^{p} m(\prod_{j=1}^{N} A_j)}$
```

Where m is the mass function, A_1 and A_2 are the two genomic sequences whereas same with B and C notations represents the subsequent fusion transcription formed in due intermediately process.

```
For each: $(A_1 \oplus A_2)$ do update( )
Compute: $m(A_1|A_2) \leftarrow \sum_{A_1 \cup A_2}^{p} m(B_1|B_2)$
Check: $m(B_1|B_2) = \frac{\sum_{B_1 \cap A_2}^{p} m(\prod_{j=1}^{N} B_j)}{\sum_{A_1 \cap B_2}^{p} m(\prod_{j=1}^{N} B_j)}$
For each $(B_1 \oplus B_2)$ do update( )
Compute: $m(B_1|B_2) \leftarrow \sum_{B_1 \cup B_2}^{p} m(C_1|C_2)$
Check: $m(C_1|C_2) = \frac{\sum_{C_1 \cap B_2}^{p} m(\prod_{j=1}^{N} C_j)}{\sum_{B_1 \cap C_2}^{p} m(\prod_{j=1}^{N} C_j)}$
For each $(C_1 \oplus C_2)$ do update( )
Compute: $m(C_1|C_2) \leftarrow \sum_{B_1 \cup B_2}^{p} m(B_1|A_2)$
Return (p,m) transcription sequence.
            end
        end
end
Stop.
```

We have modeled the epidermal growth factor receptor (EGFR) molecules as the true signal in the above algorithm which is of prime focus for intercellular communication. As shown in Fig. 1 the two membranes with direct contact binds the receptors through EGF molecules and translate the binding into specific signals of biochemical origin within the cell. Thus, triggering migration, differentiation and cell proliferation; thereby ultimately resulting in cancer or tumor. However, if this scenario of biochemical origin persists for long, it transcribes its growth in the form of biofilms as in the case of asthma and tuberculosis. These biofilms are of emergent in nature which latter develops drug resistivity as the time passes in the similar fashion as that of the redundant RNA fusion preceded by complementary and cooperative fusion based on the interions or intermediary RNA complexes formed from the fusion (Fig. 2). This dysfunction implicates in a variety of cancers, tumors, and another disease of DNA origin which retains its abnormal coding and influencing the health of the inherited offspring. Now, we have developed the TMFA algorithm for early detection of such transcription. Thereby, it is possible to develop and model a EGFR targeting drugs for irradiating such deadly viruses and bacteria which evolves with the host, while continuously degrading it. Till now the characteristic role of EGF is poorly understood in finding the cause of the diseases, but with TMFA we can detect in advance the potential threat to an individual with certain viruses and bacterial infection. However, this method requires long computational hours, yet it is successful enough to record and analyze the glycan chains attached to EGFR; which plays an essential role in the behavior of living cells.

CONCLUSION

Unlike the predecessor approach of discordantly mapping of known and paired reads annotations, TMFA - fusion can detect and distinguish both individual and paired reads that span gene fusions. These features increase its sensitivity and specificity (Fig. 3) and permit it to find fusions between novel genetic sequences and known splice variants. The TMFA effectively models the three logically different fusion states which are functionally the same from the perspective of its application but are required to be modeled as a single linkage for the sake of reproducing the results with high performance. Other methods might be used cooperatively with TMFA to operate in a distributed fashion. The one great challenges that remain out of the scope of this paper is to develop a big-data based population genome TMFA detection cycles to build an archive of potential threat of ever evolving viral and bacterial infection and simultaneously record its changing state or even may 1 day be successful in sequencing the transcription, i.e., virotherapy to eliminate such threats, in the near possible future.

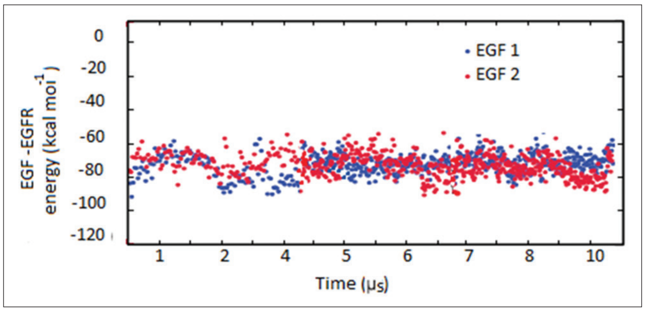

Fig. 1: Intercellular responses of two fussing RNA-sequences simulated by the type matrix fusion algorithm

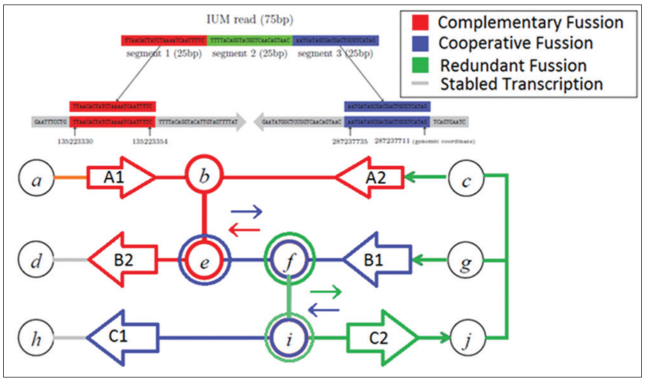

Fig. 2: Illustration of fusion types achieved from TMFA from intercellular transcription of DNA genome. Where a, d, h represents the final state whereas b, e, f, i, c, g, j are the meta states. A_1 and A_2 are the two genomic sequences whereas same with B and C notations represents the subsequent fusion transcription formed in due intermediately process

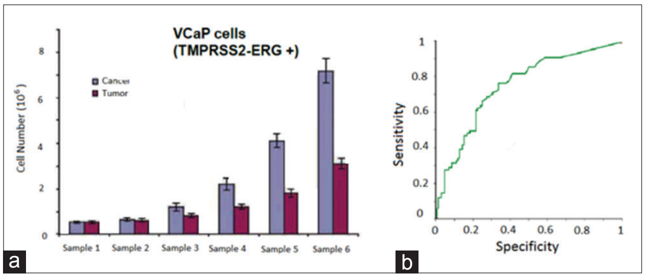

Fig. 3: (a) Sample snippet plot of six samples for the detected cancer and tumor threat from increasing number of the cell counts by the type matrix fusion algorithm (TMFA), (b) plot of the TMFA performance over 90 samples for its sensitivity versus specificity

REFERENCES

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5(7):621-8.
2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 2008;320(5881):1344-9.
3. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 2008;133(3):523-36.
4. Salzberg SL. Recent advances in RNA sequence analysis. F1000 Rep 2010;2:64.
5. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 2009;25(9):1105-11.
6. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010;28(5):511-5.
7. Rowley JD. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 1973;243(5405):290-3.
8. de Klein A, van Kessel AG, Grosveld G, Bartram CR, Hagemeijer A, Bootsma D, *et al*. A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. Nature 1982;300:765-7.
9. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, *et al.* Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458(7234):97-101.
10. Mitelman F, Johansson B, Mertens FE. Mitel man database of chromosome aberrations and gene fusions in cancer. 2011.
11. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A 2009;106(30):12353-8.
12. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 2011;12(1):R6.
13. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10(3):R25.
14. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456(7218):53-9.
15. Ameur A, Wetterbom A, Feuk L, Gyllensten U. Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol 2010;11(3):R34.
16. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, *et al.* De novo assembly and analysis of RNA-seq data. Nat Methods 2010;7(11):909-12.
17. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Res 2009;19(6):1117-23.
18. Kent WJ. BLAT - The BLAST-like alignment tool. Genome Res 2002;12(4):656-64.
19. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics 2011;27(8):1068-75.
20. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, *et al.* FusionSeq: A modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol 2010;11(10):R104.