Development of BRCA1 DNA Sequence Profile of Early Diagnosis of Breast Cancer

D. S. Nurdin[1], M. N. Isa[1], R. C. Ismail[1] and M. I. Ahmad[2]
[1]*The Integrated Circuits and Systems Design Group (ICASe),*
School of Microelectronic Engineering,
[2]*School of Computer and Communication Engineering,*
Universiti Malaysia Perlis,
Pauh Putra Campus, 02600, Arau, Perlis, Malaysia.
nazrin@unimap.edu.my

***Abstract*— Early disease detection using gene expression has been widely discussed for many diseases including breast cancer, mainly by the detection of mutations in DNA sequence. However, aligning long BRCA1 gene is rather time-consuming for early detection. The implementation of systolic architecture in FPGA which an accelerator could solve the computation time drawback. Despite that, all of the BRCA1 DNA NTs are unable to be fitted in the FPGA. Thus, the BRCA1 gene needs to be truncated for real-time detection without affecting its accuracy. The exon 11 of BRCA1 is selected as test sequence where it is truncated for 80%, 60%, 40% and 20% from 3 different locations i.e. front, middle and end. Based on the results, exon 11 which is truncated for 60% from the middle part is selected as the developed profile for early detection of breast cancer using BRCA1 gene.**

***Index Terms*— BRCA1 Gene; Breast Cancer; DNA Truncation; Pairwise DNA Sequence Alignment; SSEARCH35.**

I. INTRODUCTION

Breast cancer is a well-known disease that causes death worldwide. In Malaysia, statistics have shown that almost 25% of breast cancer death contributed to all cancer-related casualties [1]. One of the main causes of breast cancer is uncontrolled divisions of damaged cells due to the mutation of Breast Cancer Susceptibility (BRCA) gene. Breast Cancer Susceptibility Type 1 (BRCA1) and Breast Cancer Susceptibility Type 2 (BRCA2) are the two types of BRCA. Based on Figure 1, a person who inherited BRCA1 has higher chance to diagnose with both breast cancer and also ovarian cancer. Early detection of these diseases could enhance the recovery chances by finding the suitable cure. There are numerous studies on breast cancer early detection using Deoxyribonucleic Acid (DNA) sequence alignment analysis [2-4]. This method could detect Nucleotides (NTs) changes or mutations in BRCA DNA sequence.

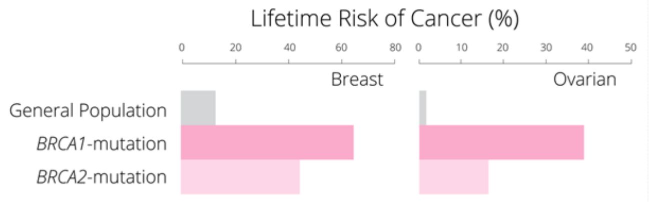


Figure 1: Lifetime risk of cancer [5]

DNA sequence alignment involves in aligning biological sequences such as DNA sequence. The fundamental of DNA sequence alignment is to search the similar NTs which are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) [6]. Pairwise sequence alignment is a type of DNA sequence alignment where it aligns a query sequence (unknown sequence) and a known subject sequence as shown in Figure 2. Mutations such as insertion and deletion could be identified during aligning DNA sequence.

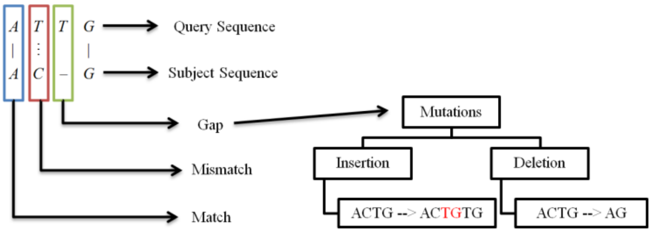


Figure 2: DNA sequence alignment.

The aforementioned method uses sequence alignment algorithm to evaluate the homology between the aligned DNA sequences. Sequence alignment algorithm can be broken down into two sub-categories; optimal and heuristic approach. The former approach searches optimal score between the aligned sequences meanwhile the latter approaches use the sub-optimal technique during alignment. The computation time and search sensitivity are the debated issues for both approaches. The optimal search is a laborious process due to its highly sensitive search. On the other hand, the heuristic search is a fast process due to its less sensitive search. However, both searches are still used until today for a better understanding of DNA sequence alignment. The Smith – Waterman (SW) algorithm [7] and the Basic Local Alignment Search Tool (BLAST) [8] are the examples of optimal and heuristic search respectively. Equation (1) is an example of the SW algorithm with affine gap penalty [9]. The optimal sequence alignment algorithm computations have been extensively discussed in [10].

Recently, the demand for advanced technology in DNA sequencing has increased due to the revolution of sequencing technology [11]. With the hand-held size of MinION [12], this research aims to develop a small device targeted for early breast cancer diagnosis that implements DNA sequence alignment method. However, it is a challenge to develop a small sequencing tool that involves long DNA sequence such as BRCA gene due to limited logic resources. In addition to this, it takes time to get a result which will result in the inefficiency of the proposed device. For example, systolic array (SA) – based architecture that implements Equation 1 was used to overcome timing problem as shown in Figure 3

[13]. With the aid of Field Programmable Gate Arrays (FPGAs) acts as a computation accelerator, sequencing all the BRCA NTs (typically 117143 NTs) in an FPGA device becomes an issue due to limited logic resources. In order to the overcome this problem, truncation of the mentioned DNA sequence is one of the best options without affecting the sequencing accuracy. Therefore, proper selection of DNA sequence in a lengthy BRCA is required.

$$M(i,j) = \max\begin{Bmatrix} I_x(i-1,j-1)+\gamma(r_i,s_j) \\ I_y(i-1,j-1)+\gamma(r_i,s_j) \\ M(i-1,j-1)+\gamma(r_i,s_j) \end{Bmatrix}$$

$$I_x(i,j) = \max\begin{Bmatrix} I_x(i-1,j)-e \\ M(i-1,j)-d-e \end{Bmatrix} \qquad (1)$$

$$I_y(i,j) = \max\begin{Bmatrix} I_y(i,j-1)-e \\ M(i,j-1)-d-e \end{Bmatrix}$$

$$F(i,j) = \max\left\{0, M(i,j), I_x(i,j), I_y(i,j)\right\}$$

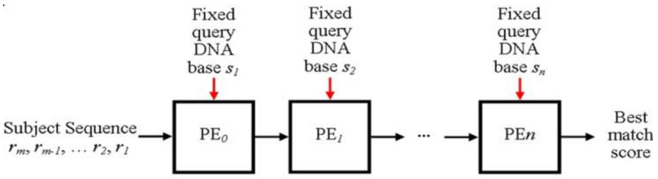
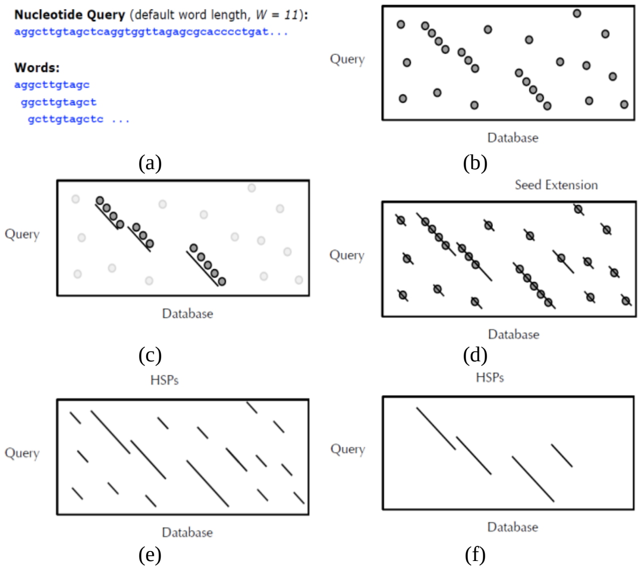

Fixed query DNA base s_1
Fixed query DNA base s_2
Fixed query DNA base s_n
Subject Sequence
$r_m, r_{m-1}, \dots r_2, r_1$
PE$_0$
PE$_1$
...
PEn
Best match score

Figure 3: Implementation of SA architecture in DNA sequence alignment [14].

The truncation idea comes from the BLAST heuristic search, where it uses local alignment from a list of high scoring words created from words similar to query sequence as shown in Figure 4. When there are similar words in the database from the word list during scanning, it starts hit extension process to extend the possible match. The words that are far from the hit extension process are eliminated.


Nucleotide Query (default word length, $W = 11$):
aggcttgtagctcaggtggttagagcgcacccctgat...
Words:
aggcttgtagc
ggcttgtagct
gcttgtagctc ...
Query
Database
Seed Extension
HSPs

(a) (b) (c) (d) (e) (f)

Figure 4: BLAST algorithm search technique [15]

In this work, exon 11 of BRCA1 DNA sequence is used to develop BRCA1 query profile as it is the largest exon among the available exons in BRCA1 with 3426 NTs. Furthermore, most of the mutations occur in exon 11 as shown in Figure 5. The truncation of exon 11 DNA sequence has been used by [16-18] for mutation detection in BRCA1 gene.

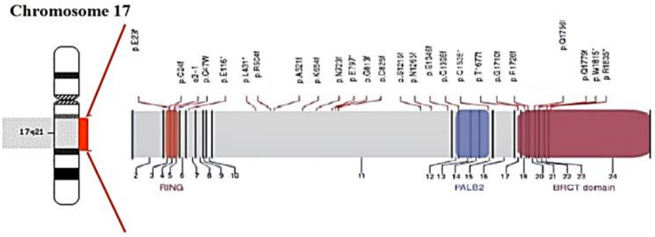


Figure 5: Exons of BRCA1 [19].

The rest of this paper is organized as follow. The following section describes the methods on developing the profile for breast cancer early detection. Section III will discuss the accuracy analysis on the sub-sequence based on truncation percentage and positions as candidates for the developed profile. Lastly, the conclusion of this research will be provided in Section IV.

II. METHODOLOGY

This section describes the method on developing the profile for breast cancer early detection by truncating the original BRCA1 sequence to four parts, i.e. 80%, 60%, 40%, and 20% for three positions as shown in Table 1 which is taken from the GenBank database (Accession No. L78833.1). Truncation locations of NTs in exon 11 DNA sequence are identified based on both aforementioned truncation percentage and positions to determine the desired test sequence as shown in Figure 6 for DNA sequencing computation.

Table 1
Various Sequence Lengths and Positions

Sequence Truncation (%)	Sequence Length	Truncation Position	NT Location
100 (Test 1)	3426	No truncation	1-3426
80 (Test 2)	2741	Front (F)	1-2741
		Middle (M)	343-3083
		End (E)	686-3426
60 (Test 3)	2055	Front (F)	1-2055
		Middle (M)	686-2740
		End (E)	1372-3426
40 (Test 4)	1370	Front (F)	1-1370
		Middle (M)	1029-2398
		End (E)	2057-3426
20 (Test 5)	685	Front (F)	1-685
		Middle (M)	1371-2055
		End (E)	2742-3426

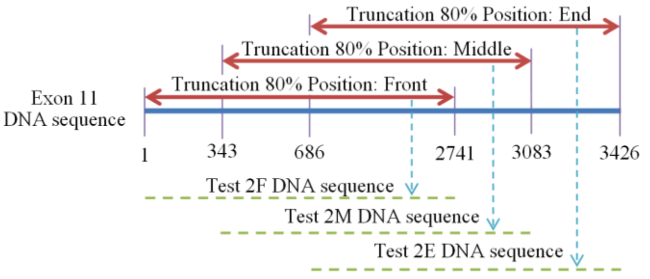
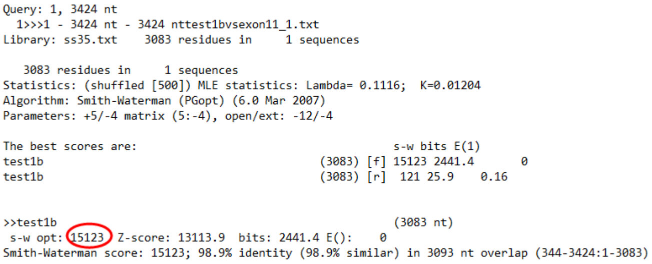


Figure 6: Exons of BRCA1 [19].

The SSEARCH35, a Fast Alignment (FASTA) package software that implements the SW algorithm with affine gap penalty as shown in Equation 1 is used for aligning DNA sequence. Based on Equation 2, the expected value is the similarity score obtained from the report that is generated by the SSEARCH35 software when the query and subject sequence both uses same test sequence as shown in Figure 7 in the red circle [20]. For example, test 1 sequence is aligned with test 1 sequence; test 2F is aligned with test 2F until test 5B. The measured value is also the similarity score obtained from the aforementioned software generated a report by aligning the test sequence and the mutated exon 11 DNA sequences as shown in Table 2. The mutated exon 11 DNA sequences are taken from the original blood sample of 11 patients where the variants were detected by CASAVA software.

$$Accuracy\ (\%) = 1 - \left(\frac{Expected\ value - Measured\ value}{Expected\ value}\right) \times 100 \qquad (1)$$


```
Query: 1, 3424 nt
  1>>>1 - 3424 nt - 3424 nttest1bvsexon11_1.txt
Library: ss35.txt    3083 residues in     1 sequences

   3083 residues in     1 sequences
Statistics: (shuffled [500]) MLE statistics: Lambda= 0.1116;  K=0.01204
Algorithm: Smith-Waterman (PGopt) (6.0 Mar 2007)
Parameters: +5/-4 matrix (5:-4), open/ext: -12/-4

The best scores are:                                      s-w bits E(1)
test1b                                         (3083) [f] 15123 2441.4      0
test1b                                         (3083) [r]  121 25.9    0.16


>>test1b                                                  (3083 nt)
 s-w opt: 15123 Z-score: 13113.9  bits: 2441.4 E():    0
Smith-Waterman score: 15123; 98.9% identity (98.9% similar) in 3093 nt overlap (344-3424:1-3083)
```

Figure 7: Report generated from the SSEARCH35 software after DNA sequence alignment analysis.

Table 2
11 Samples of Mutations in Exon 11 BRCA1 Gene [21]

Sample	Mutation
1	g37067del AA g35589A>G g35801A>G g36404T>A g36407A>T g36905A>G
2	g35269delA g36407A>T g35283C>T g36749A>G
3	g35813C>T g35512T>C g36314A>G g36529delAAGC
4	g34044delCTCA g36396T>A g36404T>A g36405T>G g36407A>T
5	g36388T>C g36389C>A g36391A>T g36396T>A g36404T>A
6	g35854delTC g36396T>A g36404T>A g36405T>G g36407A>T
7	g36388T>C g36389C>A g36391A>T g36396T>A g36404T>A g36405T>G g36407A>T
8	g35854delTC g36396T>A g36404T>A g36405T>G g36407A>T
9	g34130delA g35283C>T g35512T>C g35813C>T g36314A>G g36396T>A g36404T>A g36405T>G g36407A>T g36749A>G
10	g35154delA g35283C>T g35512T>C g35813C>T g36314A>G g36749A>G
11	g36404T>A g36405T>G g36407A>T

The accuracy of the results obtained from the experiment will be discussed in the next section with the purpose of choosing the best test sequence as the profile for breast cancer early detection.

III. Result and Discussion

As mentioned in the previous section, the test sequence is aligned with the mutated exon 11 using SSEARCH35 software for DNA sequencing process. The graph in Figure 8, shows the percentage of accuracy for test 1 (no truncation) sequence against the mutated exon 11 DNA sequence. The acceptable accuracy percentage for truncation sequence is more than 90% [22]. Based on the graph, lower accuracy shows high mutations occurs in the mutated sequence. This is because the difference between the expected value and the measured value is high. In other words, the measured value is small. The smaller the measured value, the more mutations occur in the query sequence. Please note that the fundamental of DNA sequence alignment is to find the similar NTs between the query sequence and subject sequence.

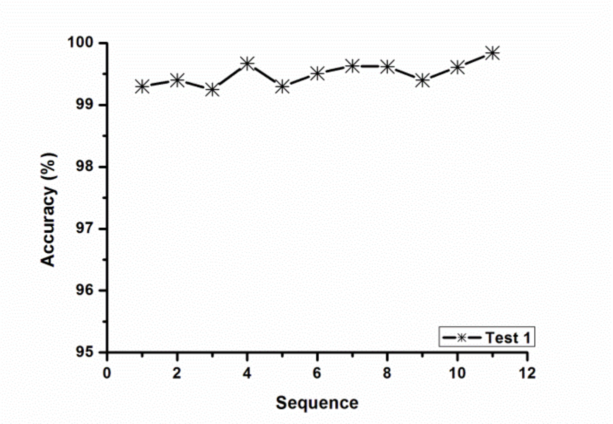
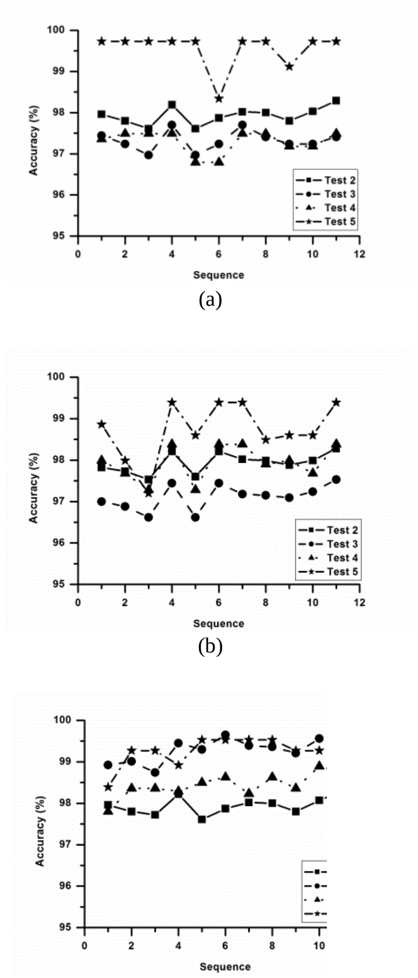


Figure 8: Graph of accuracy against sequence (mutated exon 11) for no truncation of exon 11.



(a)

(b)

(c)

Figure 9: Graphs of accuracy against sequence (mutated exon 11) for truncation positions front (a), middle (b) and end (c)

Based on the discussion, all the 12 test sequences from both different truncation positions and also percentages (query sequence) as shown in Table 1 were aligned against the mutated exon 11 DNA sequence (referred to as subject sequence) in 12 runs. According to the graph as shown in Figure 9, the middle position has the lower accuracy percentage as compared to other positions. This shows that the mutation of exon 11 is highly repetitive in middle part of the gene. Therefore, test 3 from the middle part of the DNA sequence is chosen as the query profile for early detection of breast cancer.

IV. Conclusion

In this paper, the development of DNA sequence profile for early detection of breast cancer is presented. Lower accuracy percentage shows that more mutation detected in the patients' DNA sequence. Based on the experimental results, 60% of exon DNA sequence truncation at the middle part detects more mutation as it has the lowest accuracy in average. The truncation method could overcome both the timing and also logic resources issues in hardware implementations.A conclusion to review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

Acknowledgment

This work was supported by the FRGS under Grant 9003-00480. Modeling of an Efficient DNA/Protein Sequencer using Hardware Description Language (HDL) for Biomedical Applications.

References

[1] T. M. Khan and S. A. Jacob, "Brief review of complementary and alternative medicine use among Malaysian women with breast cancer." Journal of Pharmacy Practice and Research, vol. 47, no. 2, pp. 147-152, 2017.

[2] T. Sørlie, "The Impact of Gene Expression Patterns in Breast Cancer." Clinical chemistry, vol.62, no. 8, pp. 1150-1151, 2016.

[3] D. A. Marshall, K. Deal, Y. Bombard, N. Leighl, K. V. MacDonald and M. Trudeau, "How do women trade-off benefits and risks in chemotherapy treatment decisions based on gene expression profiling for early-stage breast cancer? A discrete choice experiment." BMJ open, vol. 6, no.6, 2016.

[4] O. Balacescu, L. Balacescu, O. Virtic, S. Visan, C. Gherman, F. Drigla and O. Tudoran, "Blood genome-wide transcriptional profiles of HER2 negative breast cancers patients." Mediators of inflammation, 2016.

[5] S. M. Rosenberg, K. J. Ruddy, R. M. Tamimi, S. Gelber, L. Schapira, S. Come, V. F. Borges, B. Larsen, J. E. Garger and A. H. Partridge, "BRCA1 and BRCA2 Mutation Testing in Young Women With Breast Cancer." JAMA Oncol, vol. 2, no. 6, pp. 730-736, 2016.

[6] J. D. Watson and F. H. Crick, "The structure of DNA," in Cold Spring Harbor Symposia on Quantitative Biology, vol. 18, NY: Cold Spring Harbor Laboratory Press, 1953, pp. 123-131.

[7] T. F. Smith, and M. S. Waterman, "Identification of Common Molecular Subsequences." Journal of Molecular Biology, vol. 147, no. 1, pp. 195-197, 1981.

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool." Journal of Molecular Biology, vol. 215, pp. 403-410, 1990.

[9] G. Osamu, "An improved algorithm for matching biological sequences." Journal of Molecular Biology, vol. 162, pp. 705-708, 1982.

[10] D. S. Nurdin, M. N. Isa and S. H. Goh, "DNA sequence alignment: A review of hardware accelerators and a new core architecture." in 3rd International Conference Electronic Design (ICED) Phuket, Thailand, 2016, pp. 264-268.

[11] B. P. Hodkinson and E. A. Grice, "Next-generation sequencing: a review of technologies and tools for wound microbiome research," Advances in wound care, vol. 4, no. 1, pp. 50-58, 2015.

[12] C. Meldrum, M. A. Doyle, and R. W. Tothill, "Next-Generation Sequencing for Cancer Diagnostics: a Practical Perspective," Clin Biochem Rev, vol. 32, pp. 177 -195, Nov. 2011.

[13] .H. T. Kung and C. E. Leiserson, Systolic Arrays for (VLSI): Carnegie-Mellon University, Department of Computer Science, 1978.New York, 1994, pp. 8–16.

[14] C. Fenton, (Aug. 2009,). Final Project Report.

[15] D. H. Ardell, 'Sequence comparison, Part II: Alignments and Search', Uppsala University, 2007.

[16] A. K. Singh, A. Pandey, M. Tewari, P. Pandey, H. P. Pandey and H. S. Shukla, "BRCA1 Gene's EXON 11 and Breast Carcinoma: A Mutational Hot Spot for Familial Patients and Prone to Metastases in Northern India." InternationalJournal of Clinical and Experiment Pathology, vol. 5, no. 219, pp. 2161-2181, 2015.

[17] A. Jugessur, P. Frost, T. I. Andersen, S. Steine, A. Lindblom, A. L. Børresen-Dale and H. G. Eiken, "Enhanced detection of mutations in BRCA1 exon 11 using restriction endonuclease fingerprinting-single-strand conformation polymorphism." Journal of molecular medicine, vol.78, no. 10, pp. 580-587, 2000.

[18] A. S. Lee, G. H. Ho, P. C. Oh, C. Balram, L. L. Ooi, D. T. H. Lim and G. S. Hong, "Founder mutation in the BRCA1 gene in Malay breast cancer patients from Singapore." Human mutation, vol. 22, no. 2, p. 178, 2003.

[19] G. Liu, D. Yang, Y. Sun; I. Shmulevich, F. Xue, A. K. Soo and W. Zhang, "Differing Clinical Impact of BRCA1 and BRCA2 Mutations in Serous Ovarian Cancer." Pharmacogenomics, vol. 13, no. 13, pp. 1523-1535, 2012.

[20] W. Lichten, Data and Error Analysis. Upper Saddle River, NJ: Prentice Hall. 1999.

[21] J. D Hayward, "Next Generation Sequencing Approaches to Identify Novel Susceptibility Genes for Epithelial Ovarian Cancer" Ph. D. dissertation, University College London, 2014.

[22] M. W. Schmitt, E. J. Fox, M. J. Prindle, K. S. Reid-Bayliss, L. D. True, J. P. Radich and L. A. Loeb, "Sequencing small genomic targets with high efficiency and extreme accuracy." Nature methods, vol. 12, no. 5, pp. 423-425, 2015.