Asian Journal of Information Technology 15 (1): 76-81, 2016
ISSN: 1682-3915

Micro Sequence Identification of Bioinformatics Data Using Pattern Mining Techniques in FPGA Hardware Implementation

[1]A. Surendar, [2]M. Arun and [3]A.M. Basha
[1]Anna University, Chennai, India
[2]School of Electronics Engineering, VIT University, Vellore, India
[3]Department of ECE, KSR College of Engineering, Tiruchengode, India

Abstract: The growing importance of medical solutions requires special hardware and special devices to perform high dimensional data processing. The medical problems like gene selection, protein sequence identification and DNA sequence detection has great impact in this area. To perform such high dimensional process, it requires special hardware implementation and designing such implementation also increases the complexity of efficiency. The FPGA (Field Programmable Gate Arrays) is the well-known design and we propose a novel algorithm for sequence detection in any of bioinformatics data. Unlike previous methods, the proposed method identifies the sequence of each factor on the basis of their occurrence. For each size of sequence, the method performs matching to find out the similarity of the sequence. Each of sequence is named as a pattern and based on the pattern being identified the method computes the similarity between each of the samples available. The method computes the multi level similarity measure with available sequences. Based on the multi level similarity measure computed a single sequence of bioinformatics can be identified. The proposed method produces efficient result in sequence detection and improves the hardware utilization and reduces the time complexity as well.

Key words: Bioinformatics, pattern matching, FPGA implementation, micro sequence identification

INTRODUCTION

The bioinformatics is the information which represents the biological species of human. For example, the DNA sequence of any human can be presented as informatics and has many information and similarly, the protein sequence is also a biological information of any human (Eddy, 2002). However, the DNA sequence of any human is upto a million and in many situation the researchers are challenged in identifying the similar DNA sequences. Identifying similar sequence in high dimensional data is not such easier and requires some special attention.

People with similar sequence of DNA species have similar habits and other features. So, in order to classify a person according to his habits or based on some other factor, the DNA sequences could be used. Similarly, the classification of DNA sequence could be used to perform various activities. Not only could this but the protein sequence be used by different organization in identifying the exact food type based on the result of protein classification. However, identifying such similar sequences is quite challenging task and to perform such task will take more time.

In medical domain, the requirements of fast efficient and accurate solutions are increasing due to the size of bioinformatics as well as the amount of information available. To perform such task, the researchers have invented various architectures among them the popular field programmable gate array is the efficient one which produces efficient results. The FPGA implementation is capable of computing large dimensional input and produce small set of outputs. Also, the FPGA architecture is capable of handling varying size of input values and works according to the size of input size.

This study works over the FPGA implementation and how the efficiency of FPGA can be adapted to the problem of bioinformatics problems. We focus on the micro sequence identification of DNA sequences and amino acid sequences. Regarding the problem of sequence identification (Sujithra and Padmavathi, 2015), the pattern matching solution can be adapted and capable of producing efficient results. The pattern matching is the process of identifying the presence of similar sequence in the DNA sequence or amino acid sequence. For example, the amino acids can be classified into three groups namely α-A, β-B and γ-G. The Alpha group covers the acids namely G, H, I and the Beta group covers the acids namely B and E. The gamma group covers the remaining acids from the acid sequence. For the given acid sequence, "MEKLL DEVLAP GGPYNL TVGS WVRDH VRSIV EGA WEVR", the pattern can be generate

Corresponding Author: A. Surendar, Anna University, Chennai, India

Table 1: Example patteren generated

Amino acid sequences							
1	2	3	4	5	6	7	8
CB	CC	CCC	CCCC	CBCCCB	CCAACCCC	CCACCACCBCC	CCCCACCCCCACCA
CC	CC	CCC	ACCC	CCCCAA	CCACCCCC	CBCCCBCCCCAA	CBCCCBCCCCAACCC
BC	AC	ACC	CACC	CCCCCC	ACCACCBC	CCCCCCACCCCC	CCCACCCCCACCACC
CC	CA	CCA	ACCB	ACCCCC	CBCCCBCCC	CBCCCBCCCCAA	CBCCCBCCCCAACCCC
CC	CC	CCA	ACCA	ACCACC	CAACCCCCC	CCCCCCACCCCC	CCACCCCCACCACCBC
AA	BC	CCB	CBCCC	CBCCCBC	ACCCCCACC	CBCCCBCCCCAA	CBCCCBCCCCAACCCCC
CC	CBC	BCC	BCCCC	CCCAACC	ACCCCCCACCC	CCCCCCACCCCC	CBCCCBCCCCAACCCCCC
CC	CCB	CBCC	AACCC	CCCCACC	CBCCCBCCCC	CBCCCBCCCCAA	CBCCCBCCCCAACCCCCCA

as follows: CBC CCB CCC CAA CCCC CCA CCCC CAC CAC CBCC, which is the complete pattern. In order to find the sequence matching from the large set of amino acid sequence the pattern mining technique can be used. Different acid sequence would have different pattern and for any group of sequence we can identify similar pattern (Arun and Krishnan, 2012). The size of sequence matching can be computed by identifying different size of patterns.

Micro sequence identification is to identify the sequence matching in different level for example from two set sequence to the maximum size. The acid sequence may match at different number of levels and based on the matching in different sequence levels the similar sequence can be identified and used for classification. Such multi level sequence identification can be used for DNA classification and for other bioinformatics solutions.

Literature review: There are number of methods has been described earlier for the problem of sequence identification and we discuss some of the method here in this study.

The Needleman-Wunsch (NW) algorithm (Muhamad *et al*., 2015) performs an optimal global alignment of two sequences based on certain constraints. The algorithm aligns the sequences based on maximum number of matches in amino acid and minimum number of gaps required to align the sequences. It is sometimes referred to as the optimal matching algorithm, because the Needleman-Wunsch algorithm (Muhamad *et al*., 2015) finds the optimal alignment of the entire sequence of both proteins, it is a global alignment technique. Consider two strings of gene characteristics s and t where s is ATTGCTCTG and t is ATGCCG. In these sequences of varying lengths as given in Table 1, it can be found by introducing few gaps; the maximum alignment of the sequences can be maximized. The score of alignment of two elements is 1. If there is a mismatch or gap in the resultant of the algorithm, the score is -1. The cost of entire alignment would be the sum of all scores for the alignment.

The Smith-Waterman algorithm, dynamic programming algorithm, is for determining analogous regions between two nucleotide or protein sequences. It is a non-heuristic algorithm that guarantees to find the optimal local alignment with respect to the scoring system being used in Needleman-Wunsch algorithm. Smith-Waterman algorithm (Muhamad *et al*., 2015; Khajeh-Saeed *et al*., 2010) reduces the number of counterfeit positives.

However, the Smith-Waterman algorithm produces optimal results by demanding of time and memory resources and by sacrificing speed. To confirm a solution to be accurate and optimal, Smith-Waterman is required to be more computations costing its reputation to taken the position of a fast algorithm. As a result, it has been reinstated in by its successor, BLAST algorithm with an option of approximate solution.

The blast algorithm (Suseela, 2014) takes the input as the genetic sequence database and a query which has to be found in the database. The outputs of the algorithm are the positions of the areas of these two strings that have similarity, as well the score of these similarities. The quality of each pair-wise alignment (Surendar *et al*., 2015) is represented as a score and the scores are ranked. Scoring matrices are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein). The alignment score will be the sum of the scores for each position. The significance of each alignment is computed as E-value. The lower the E-value (Altschul *et al*., 1997), the more significant is the score and the sequences are homologues for low E values. Each of these pairs, comprising of a database area and a query area, is called a High Score Pair (HSP). The score has significantvalue for biologist because it is used to compute several variables, of which the E-value is the most important.

FPGA based agrep for DNA microarray sequence searching utilizes the capability of FPGA to parallelize such processes and introduces a hardware-based implementation of Agrep, a fast text searching algorithm capable to allow approximate matches. The design was implemented in Opal Kelly® XEM3010 and was tested using DNA microarray sequences from the NCBI virus probe database. Results indicate significant improvement in performance in terms of runtime and throughput as

compared to a software-based Agrep (Pearson and Lipman, 1988; Lavenier, 2009). Low power bloom filter architectures using multi stage lookup techniques (Khajeh-Saeed *et al*., 2010) introduce a low power bloom filter architecture, which is space and power effective in hardware platforms. Instead of working on programming phase or technology, our work concentrates on lookup techniques of bloom filters (Surendar *et al*., 2013; Pearson and Lipman, 1988). In this study, we propose a new multi-stage lookup technique for bloom filters and the theoretical power analysis of the proposed lookup techniques is presented. Power analysis shows that a decrease in the number of hash functions per stage results in power gain.

All methods has the problem of time complexity and power overhead in the sequence identification.

MATERIALS AND METHODS

Micro sequence identification using pattern mining: The micro sequence identification approach reads the input sequence and for each class of amino acid sequence, the method generates number of pattern from 2 to N, where N is the size of sequence. For each class according to the patterns being generated, the method computes the Multi Level Sequence Similarity Measure (MLSSM) which represents how far the input sequence is close to the sequences belongs to the different classes. The entire process can be split into number stages namely multi level pattern generation, multi level sequence similarity measure computation, sequence identification.

Figure 1 shows the architecture of multi level micro sequence identification and shows the components of the proposed approach.

Multi-level pattern generation: The multi level pattern is the combination of different pattern for different dimension. For given DNA sequence or amino acid sequence with dimension D, the method generates N number of D_n dimensional sequence from 2 to D. For each class of DNA sequence, the method generates number of different sequence pattern at different levels. The pattern is generated from different dimension and generated pattern will be used to compute the sequence similarity measure.

Algorithm A: Generates DNA pattern

Pseudo Code of MLP Generation:
Input: DNA Sequence Ds
Output: Pattern Set Ps.
Start
    Compute the size of sequence ss.
    Ss = ∫ size (Ds)
    For each dimension Di from SS
    Generate Pattern Pi.
    Pi = $\int^{SS}_{I=1}$ subset (Ds,Di)
        Add to pattern set Ps.
    Ps = ∑ (pk ∈ ps) u pi
    End
Stop

The algorithm generates the multi level pattern set from given DNA sequence and generated sequence will be used to compute the sequence similarity measure.

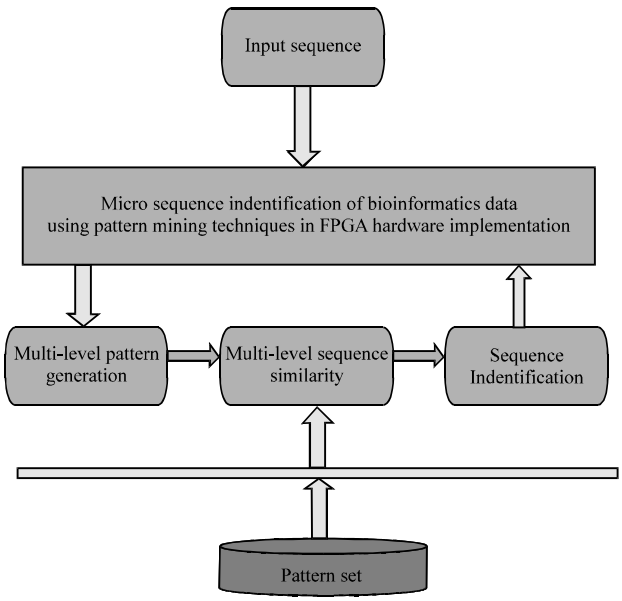

Fig. 1: Architecture of multi level micro sequence

CBCCCBCCCCAACCCCCCACCCC
CBCCCBCCCCAACCCCCCACCCCC
CBCCCBCCCCAACCCCCCACCCCCA
CBCCCBCCCCAACCCCCCACCCCCAC
CBCCCBCCCCAACCCCCCACCCCCACC
CBCCCBCCCCAACCCCCCACCCCCACCA
CBCCCBCCCCAACCCCCCACCCCCACCAC
CBCCCBCCCCAACCCCCCACCCCCACCACC
CBCCCBCCCCAACCCCCCACCCCCACCACCBCC

Fig. 2: Example pattern generated

For example, from the given amino acid sequence "MEKLLDEVLAPGGPYNLTVGSWVRDHVRSIVEGAW EVR", the pattern generation approach can generate the following patterns as follows:

According to the representation and the class of amino acids the input sequence is represented as follows: "CBCCCBCCCCAACCCCCCACCCCCACCA CCBCC".

Table 1 shows the set of patterns being generated from the above discussed algorithm. Figure 2 shows the example pattern being generated by the proposed algorithm.

Table 1 and Fig. 2, shows the set of patterns being identified from the starting position and similarly the patterns can be generated from the remaining dimensions which produce enormous number of patterns. The method generates such patterns and will be used to compute the sequence similarity measure.

Multi-level sequence similarity measure: The multi level sequence similarity measure shows the similarity of the sequences at different levels and the number of levels is depending on the dimension of the sequence. For each level using the pattern set being generated, the method computes the sequence similarity measure. For each dimension the method computes the similarity measure and then finally, the method computes the multi level similarity measure which will be used to identify the sequence.

Algorithm B: Complete the multiple level

Pseudo Code of MLSSM:
Input: Pattern Set Ps, Pattern Pi
Output: MLSSM.
Start
For each level l from Ps
Compute similarity.
MLSSM = $\int_{i=1}^{Levels}\int_{j=1}^{size(Ps)}\sum Ps(j,l)==Pi$
End
MLSSM = MLSSM/size (levels)
Stop

The algorithm B computes the multi level sequence similarity at each level and finally computes the cumulative sequence similarity.

Sequence identification: At this stage, the method generates the sequence set for the given sequence and based on the pattern set being generated the method computes the multi level sequence similarity. For each level of the DNA sequence, the method computes the sequence similarity and finally, the method computes the cumulative sequence similarity measure. For each class, the method maintains different sequence and the method computes multi level sequence similarity for each class. Based on the sequence similarity measure the method selects a single class and identifies the most sequence similar.

Algorithm C: Computes multi-level sequence

Pseudo Code of Sequence Identification:
Input: DNA sequence Ds, Pattern Set Ps
Output: Sequence S.
Start
Pattern set Dps = Multi Level Pattern Generation(Ds).
For each class Ci
For each Level l
Compute MLSSM.
MLSSM = $\int_{i=1}^{size(DPs)}\sum MLSSM(Ps,DPs(i))$
End
MLSSM = ΣMLSSMi/size (Ps)
End
Choose the class with maximum MLSSM.
Choose the sequence with maximum MLSSM.
Stop

The algorithm C computes the multi level sequence similarity measure and selects the class and sequence with maximum sequence similarity measure.

RESULTS AND DISCUSSION

The proposed multi level micro sequence has been implemented and evaluated for its efficiency using the model simulator and has been evaluated using the FPGA test bench. The method has been validated for its efficiency using various DNA sequence and amino acid sequence. The efficiency of the method has been validated by computing the sequence detection accuracy and the time complexity produced.

Table 2 shows the details of data set being used to evaluate the performance of the proposed approach. The method has been validated for its efficiency using different data sets and the method has been validated for its efficiency in sequence identification and its time complexity.

Figure 3 shows the comparison of sequence identification efficiency produced by different methods

Table 2: Details of data set used

Data set	Size
GENIE	793
UCI	2500
dbGap	4300

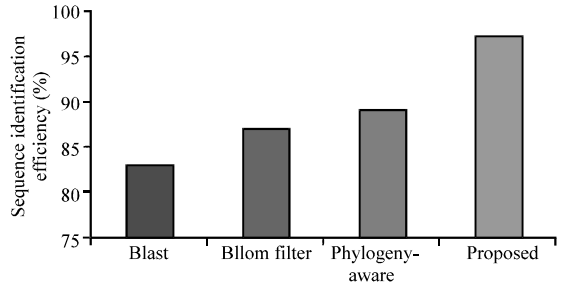

Fig. 3: Comparison of sequence identification efficiency

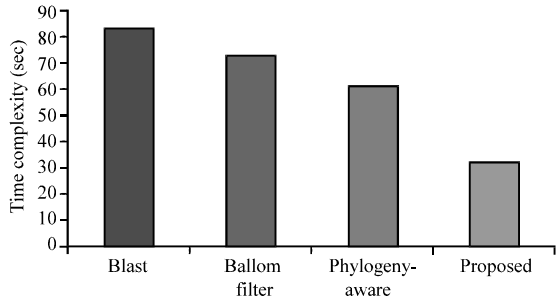

Fig. 4: Comparison of time complexity

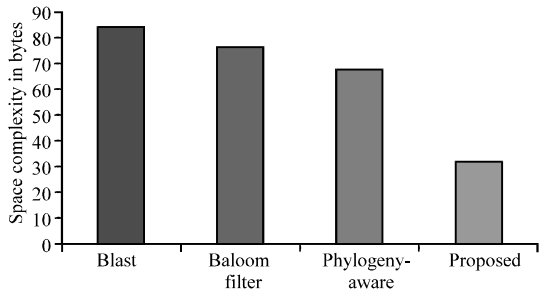

Fig. 5: Comparison of space complexity

and it shows that the proposed method has produces higher efficiency than other methods. Figure 4 shows the comparison of time complexity produced by different methods in identifying the sequence and the result shows that the proposed method has produced less time complexity than other methods.

Figure 5 shows the comparison of space occupied by the different methods and it shows that the proposed method has produced less space complexity than other methods.

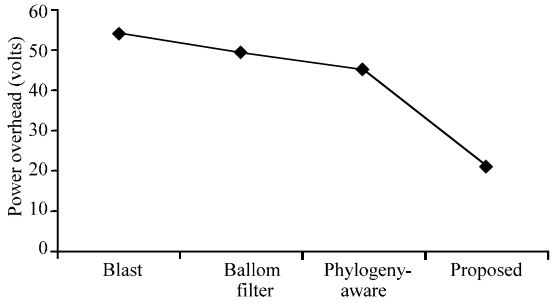

Fig. 6: Comparison of power overhead

Figure 6 shows the comparative results on power overhead produced by different methods and the result shows that the proposed method has produced less power overhead than other methods.

CONCLUSION

In this study, we propose a micro sequence identification using pattern mining technique. First, the method generates number of patterns or sequences from the dimension 2 to the dimension N. The patterns are generated at each dimension and with varying size of dimension. The generated patterns are stored in the pattern set and for the input sequence the method generates the similar set of pattern set. Based on generated pattern set, the method computes the sequence similarity at each level and finally, a cumulative similarity value is computed. Based on the value of multi level sequence similarity value, the method selects the sequence as the result. The proposed method identifies the DNA sequence in efficient manner and reduces the time complexity.

REFERENCES

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang and Z. Zhang *et al.*, 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl. Acids Res., 25: 3389-3402.

Arun, M. and A. Krishnan, 2012. Functional verification of signature detection architectures for high speed network applications. Int. J. Autom. Comput., 9: 395-402.

Eddy, S.R., 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics, Vol. 3. 10.1186/1471-2105-3-18

Khajeh-Saeed, A., S. Poole and J.B. Perot, 2010. Acceleration of the Smith-Waterman algorithm using single and multiple graphics processors. J. Comput. Phys., 229: 4247-4258.

Lavenier, D., 2009. PLAST: Parallel local alignment search tool for database comparison. BMC. Bioinf., Vol. 10 10.1186/1471-2105-10-329

Muhamad, F.N., R.B. Ahmad, S.M. Asi and M.N. Murad, 2015. Reducing the search space and time complexity of needleman-wunsch algorithm (global alignment) and Smith-waterman algorithm (local alignment) for DNA sequence alignment. J. Technol., 77: 137-146.

Pearson, W.R. and D.J. Lipman, 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA., 85: 2444-2448.

Sujithra, M. and G. Padmavathi, 2015. A survey of biometric Iris recognition: Security, techniques and metrics. Asian J. Inf. Technol., 14: 192-199.

Surendar, A., M. Arun and C. Bagavathi, 2013. Evolution of reconfigurable based algorithms for bioinformatics applications: An investigation. Int. J. Life Sci. Bt. Pharm. Res., 2: 17-27.

Surendar, A., M. Arun and P.S. Periasamy, 2015. A parallel reconfigurable platform for efficient sequence alignment. Afr. J. Biotechnol., 13: 3344-3351.

Suseela, R.V., 2014. An efficient retouched bloom filter based word-matching stage of blastn. Int. J. Eng. Sci. Res., 1: 25-31.