Efficient Algorithm Implementation for Processing Large Scale Proteomics Data

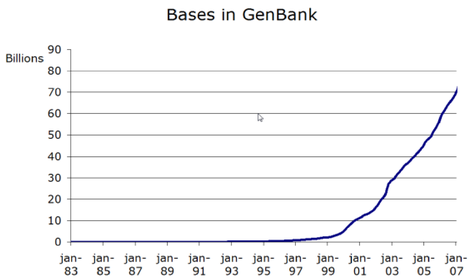
S.M. Vidanagamachchi[1] , S.D Dewasurendra[1], R.G. Ragel[1] and M.Niranjan[2]

[1]Department of Computer Engineering, University of Peradeniya, [2] School of Electronics and Computer Science, University of southampton, UK.

ABSTRACT

The concept of proteome is introduced recently and it can be used for protein identification. It requires peptide sequencing first. Then we have to use efficient computational methods to match peptides within proteins in order to identify unknown proteins. Because of the unprecedented rate of data growth, the need for fast, reliable sequence comparison engines is growing. Since software implementations consuming more time than parallelized, application specific hardware implementations, the current trend is to use hardware/software solutions such as Field Programmable Gate Array (FPGA) implementation.

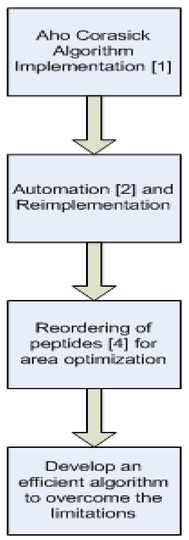
MOTIVATION



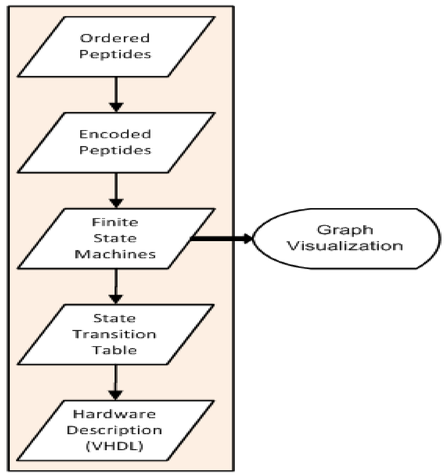
Due to advances in life sciences, there is a rapid growth of biological data including DNA sequences, protein sequences and gene expression data. Therefore, there is a pressing need for efficient computational methods to analyze and process them. However, a significant bottleneck exists in the analysis of such data. Computational demand for analyzing huge amount of biological data is growing faster than the increase in processing power of computers. Many attempts have been made in the past to develop efficient algorithms as well as dedicated hardware/software solutions to deal with this explosion.

METHODOLOGY

There are a number of hardware/software implementations of some string matching algorithms (both approximate and exact) in the past. Since Aho-Corasick algorithm is the best and the widest used multiple pattern matching algorithm which searches all occurrences of any of a finite number of keywords in a text string, Yoginder et al. [1] have used this algorithm for hardware acceleration of peptides pattern matching for the 1st chromosome of human genome. We have re-implemented and automated [2] the work done by Yoginder et al. with different data sets on a different FPGA (Altera Cyclone II FPGA) from the one they used (Xilinx). They have used bit split implementation of Aho-Corasick to reduce storage space. Resource utilization and processing times were compared for different peptide lengths [3]. However there is a hardware limitation there; we couldn't reconfigure our FPGA for 2800 peptides. Therefore we have done the area optimization in FPGA for peptide identification to map more peptides to a tile (small logic area in FPGA). Then a peptide reordering performed in order to optimize the tile utilization is of the peptides [4].

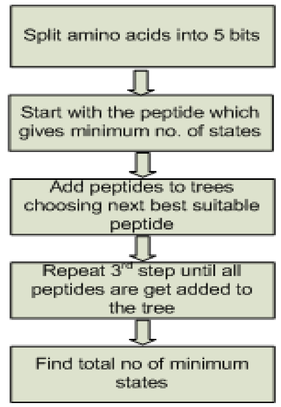


AUTOMATION

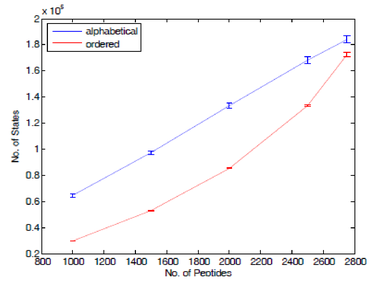


Here we have shown the whole process of automation. Initially we give the reordered peptide set to the system and then it generates the state transition table, VHDL(VHSIC Hardware Description Language) code and graph visualization automatically.

REORDERING PROCESS



RESULTS



FUTURE WORK

According to our reordering method we need to figure out the best suitable types of biological data (considering length) that gives better results for our optimization.

Our final aim is to come up with an efficient multiple pattern matching algorithm and develop it in hardware. We have studied several limitations of existing multi-pattern matching algorithms and most of them were not implemented in hardware. Aho Corasick algorithm is the only one that is implemented on FPGA.

REFERENCES

[1] Yoginder S. Dandass, Shane C. Burgess, Mark Lawrence and Susan M. Bridges, Accelerating String Set Matching in FPGA Hardware for Bioinformatics Research, BMC Bioinformatics, vol. 9, pp. 1–11 2008

[2] S.M. Vidanagamachchi, S.D. Dewasurendra, R.G.Ragel and M. Niranjan, Automated Efficient Method for String Matching using FPGA, Peradeniya University Research Sessions, Vol. 15, pp 627-629 ,2010

[3] S.M. Vidanagamachchi, S.D. Dewasurendra, R.G.Ragel and M. Niranjan, Introduction to Basic Genetics and Hardware Acceleration of Protein Sequence Data Processing, Peradeniya University Research Sessions, Vol. 15, pp 630-632 , 2010

[4] S.M. Vidanagamachchi, S.D. Dewasurendra, R.G.Ragel and M. Niranjan, Tile Optimization for Area in FPGA based Hardware Acceleration of Peptide Identification, Sixth International Conference on Industrial and Information Systems, pp 140-145, 2011