Introduction

Identifying the new genes and predicting their functionality is the common task in the bioinformatics. If the first step is straightforward (relatively), the prediction one requires algorithms to achieve a goal. The well - known way of doing that is sequence alignment, which is separate topic of bioinformatics [1]. The working principle of the sequence alignment is a comparison of two inputs, where one of them is query input and other one is DNA database [2].

It has two types regarding the way of determining the function of genes. The first one is global alignment, which is not practical in real life, that compares two long sequences and will give a negative result even when there are strong small DNA sequence matching. The local alignment is the useful and applicable type, which takes small DNA query and assimilates it through database.

According to the above mentioned type there are some algorithms with the aim to reduce computation steps and time. Smith Waterman [3], FASTA [4] and BLAST are the famous methods for the sequence alignment. Smith Waterman is the exhaustive algorithm that has the proper output but with bug called time. Another methods (FASTA and BLAST) use heuristic way of computation to overcome an issue with steps and time, also, they have optimal values as an output.

Background

FPGAs (Field Programmable Gate Array Logic) are widely applied in sequence comparison in Large Sequence databanks for the reason that they significantly improve performance due to parallel architecture. There are number of proposed and applied implementations of BLAST algorithm on FPGA in order to address the issue of biological sequence alignment [\*1]. Examples of proposed hardware implementations can be Mercury BLASTn [\*2], Mercury BLASTp [\*4], Tree-BLAST [\*5], RC-BLAST [\*6], FPGA/FLASH Accelerator [\*3], Multiengine BLASTn Accelerator [\*7].

The architectures mentioned above are based on the Word Position Record-Based Search (WPRBS) method[\*8], which stores words of query sequence in a constructed storage table and compares the subject sequence with storage table to detect the hit. The strategy is widely applied; however performance and searching capacity limitations are still extant. The reason is that during one clock cycle only one word can be searched, and thus more than one hit per clock cycle cannot be detected. Timing issue is also caused by shortage in WPRBS searching capacity due to limited number of memory ports in FPGAs.

Examples of such architecture are Mercury BLASTn [\*2] and Mitrion [\*9]. In these two systems storage tables are stored in the external memory SRAM that is attached to the FPGA, since the tables storing the words of long query sequences will require vast amount of space in internal memory RAM. Searching is performed by comparing one subject sequence from database to every word in the storage table simultaneously during one clock cycle. Apparently, performance of the system will be low: accessing external memory SRAM for getting subject sequence from database every time consumes significant amount of time.

Performance could be improved by FPGA/FLASH [\*3], which minimizes accesses to the external memory by enabling to detect several exact hits simultaneously. It is due to index structure of the database: words in the sequence associated with their position in the database and neighboring elements. However, huge space in the memory is needed to store index of the database as well as database itself. For database exponentially growing every year this architecture will not be efficient.

In order to solve the timing isuse, the architecture Multiengines BLASTn [\*7] is constructed in such way that compares 64 subject sequences simultaneously due to its 64 identical computing units. The architectures described above are based on WPRBS.

There is another method based on systolic array that does not construct tables. Needleman-Wunsch algorithm [\*10] and dynamic programming algorithm that apply systolic array was implemented on SPLASH2 by D. Hoang et. al. [\*11], [\*12]. Another algorithm that uses systolic array called Smith Waterman matching algorithm was implemented by S. Guccione *et. al.* [\*13], also at Virginia Tech [\*14] and Nanyang Techological University [\*15].

BLAST Algorithm

The Basic Local Alignment Search Tool (BLAST) is the most widely used algorithm for sequence alignment. The reason is due to the open source software National Center for Biotechnology Information (NCBI) that provides a huge DNA database for research purposes (not only).

Depending on the combination of protein and amino-acid the BLAST has four general types (Fig 1)

|  |  |  |  |
| --- | --- | --- | --- |
| Database | Query | | |
|  | DNA | Protein |
| DNA | BLASTn | TBLASTn |
| Protein | BLASTx | BLASTp |

*Fig.1*. Types of the BLAST

As the name suggests it uses the local alignment type and in general contains three steps to give the result. Also, it is important to mention that the output of the BLAST will be as score with location in the database.

i. Word Search Method. The BLAST will divide the input query into the “words” with length k, where k can be three or eleven depending on the type (see section 2.1) The words will be as shown below, following words will shift by one to the right. “ATGGGCCAGGA” is the first word and “CTTTGGGAAA” is the last one as an example.



ii. Search Database “hits”. It will find the areas where there is a matching between input query from 1st step and the DNA database.

iii. Maximum High Score Pair (HSP) . Extend the matching area till it reaches its maximum score.

Reference list:

[1]<http://www.ks.uiuc.edu/Training/Tutorials/science/bioinformatics-tutorial/bioinfo rmatics.pdf>

[2]<https://www.ncbi.nlm.nih.gov/>

[3]Optimization Strategies for Smith-Waterman Algorithm on FPGA Platform [4] Fast Dynamic Programming Based Sequence Alignment Algorithm