A New Hardware Architecture for Genomic and Proteomic Sequence Alignment

Greg Knowles and Paul Gardner-Stephen
Embedded Systems Laboratory,
School of Informatics and Engineering, Flinders University,
GPO Box 2100, Adelaide 5001, Australia
gknowles@infoeng.flinders.edu.au gardners@infoeng.flinders.edu.au

Abstract

We describe a novel hardware architecture for genomic and proteomic sequence alignment which achieves a speed-up of two to three orders of magnitude over Smith-Waterman dynamic programming (DP) in hardware [1]-[7]. In [8, 9] we introduce several features of our search algorithm, DASH, which outperforms NCBI-Blast (BLAST) [10] by an order of magnitude in software, and has better sensitivity. Indeed, DASH has been shown to have excellent sensitivity compared to Smith-Waterman. It is designed around the principle of considering genomic and proteomic sequence alignments to typically consist of regions of high homology (the diagonals) interspersed with regions of low homology. In DASH, the optimal solution consists of such diagonals joined by regions of exact DP. This is affordable due to the small area of these inter-connecting regions, Accordingly, we have designed a chip which finds the diagonals and performs the inter-region DP directly in hardware. On a Xilinx Virtex II, XC2V6000, FPGA, it performs over 10^{12} *base comparisons/second.*

1. Introduction

Heuristic algorithms such as BLAST [10] are indispensable for searching todays large genomic and proteomic databases. A significant contributor to the search time for such algorithms is the DP evaluations, BLAST spends around 76% of its time budget in this area. We have developed a novel sequence alignment algorithm, DASH, which we have shown to be superior to BLAST in both speed and sensitivity [8, 9], and excellent sensitivity with resoect to Smith-Waterman. Our project was originally motivated by the desire to develop an algorithm which would allow extensive parallelism for optimum use of reconfigurable hardware, such as FPGAs. In this context we designed DASH around the principle of considering genomic and proteomic sequence alignments to typically consist of regions of high homology interspersed with regions of low homology, Figure 1. Finding the regions of high homology (the diagonals, zone 1 in the figure) requires high throughput, but low processing power, which is ideally matched to the massively parallel resources available on FPGAs (Field Programmable Gate Arrays).

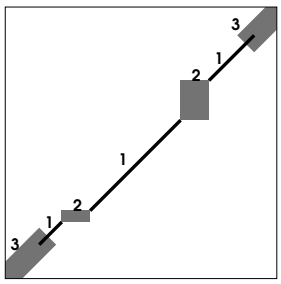

Figure 1. Illustration of the DASH Algorithm and DP Zones

The regions of low homology, zones 2 and 3 in the figure, are solved by DP, but as they are strictly limited in size, the extra hardware resources required are small.

2. The Architecture

The architecture of our design is given in Figure 2. The query and subject sequences are input to a match pipeline which finds the diagonals, Figure 3. Here, each individual match unit finds either four base matches in the nucleotide case, or one amino acid match in the protein case. Substitutions, but not gaps, are allowed. Those diagonals of more than a specified length and score are then output to a FIFO (first-in first-out buffer). The DP units then take the diag-

Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)
0-7695-2194-0/04 $20.00 © 2004 **IEEE**

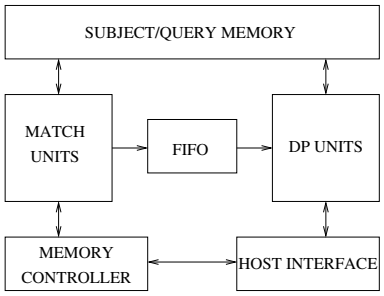

Figure 2. Architecture of the Sequence Alignment Hardware

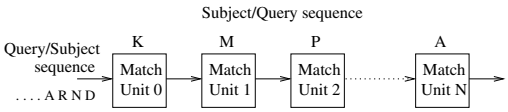

Figure 3. Match Unit Pipeline

onals from this FIFO and find the optimum path for the inter-diagonal regions, zone 2, and the ends, zone 3. The raw data is filtered by the Match Unit at high-speed (Table 1) giving the DP Unit more than enough time to process the inter-diagonal regions. Given sufficient on-board memory both these units can operate in parallel, and the critical overall system delay is the throughput of the Match Unit. There are several well-known architectures for the processing elements in the DP Unit, e.g. [1]-[6].

3. Results and Conclusions

The match units were coded in VHDL and extensively tested against DASH with DNA/Protein sequences of up to half a million in length. The area (no of slices) and timing results for a single match unit after synthesis and place and route are given in Table 1. It can be seen that in the nucleotide case over $10,000$ match units will fit on a Xilinx Virtex 2, XC2V6000, FPGA (total of 33792 slices), with a maximum throughput of over $2x10^{12}$ base comparisons/sec. With only 50% utilization of the device, we still obtain over 10^{12} base comparisons/sec, and allocate the other half of the device to the DP Unit, memory controllers, and the interface unit for the host processor. The maximum throughput in the protein case is $3x10^{11}$ amino acid matches/sec.

Results for full DP for the same FPGA can be found in [5]. Depending on the maximum query length, they obtain a peak throughput of $5.7x10^{9}$ base matches/sec. Clearly, the architecture described here affords two to three orders of magnitude greater performance.

	DNA	Protein
Slices	3.2	19.2
Clock	210 MHz	180 MHz

Figure 4. Area and Timing results for a DNA/Protein Match Unit

4. Acknowledgements

We would like to thank the Australian Co-Operative Research Center for Sensor, Signal and Information Processing (CSSIP), project Firmware for Genomics, for supporting this research.

References

[1] D.T Hoang, "Searching Genetic Databases on Splash 2", *Proc. IEEE Workshop on FPGAs for Custom Computing*, IEEE CS, 1993, pp. 185-191.

[2] P. Guerdoux-Jamet, D. Lavenier, "SAMBA:Hardware Accelerator for Biological Sequence Comparison", *CABIOS* 12(6), pp. 609-615.

[3] S.A. Guccione, E. Keller, "Gene Matching using JBITS", *Proc. 12th Int. Workshop on Field-Programmable Logic and Applications (FPL'99)*, Springer, LNCS 2438, 2002, pp. 1168-1171.

[4] B. Schmidt, H. Schroeder, M. Schimmler, "Massively Parallel Solutions for Molecular Sequence Analysis", *Proc. 1st IEEE Int. Workshop on High Performance Computational Biology*, Ft. Lauderdale, Florida, 2002.

[5] T. Oliver, B. Schmidt, "High Performance Biosequence Database Scanning on Reconfigurable Platforms", Preprint, 2004.

[6] C.W. Yu, K.H. Kwong, K.H. Lee, P.H.W. Leong, "A Smith-Waterman Systolic Cell", *Proc. 13th Workshop on Field Programmable Logic and Applications (FPL'03)*, Springer, LNCS 2778, 2003, pp. 375-384.

[7] TimeLogic Corporation, `http://www.timelogic.com`

[8] P. Gardner-Stephen, G. Knowles, "DASH: A New High Speed Genomic Search and Alignment Tool", *4th International Conference on Mathematics and Computers in Biology and Chemistry (MCBC 03)*,1, 2003, pp. 121-127.

[9] P. Gardner-Stephen, G. Knowles, "DASH: Localising Dynamic Programming for Order of Magnitude Faster, Accurate Sequence Alignment", *3rd IEEE Computational Systems Bioinformatics Conference*, Stanford, 2003.

[10] Stephen F. Altschol et al, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25, 1997, pp. 3389-3402.