CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment

ID - 171

Manavski Svetlin[1], Mariano Antonio[1], Valle Giorgio[2]
[1]Elaide srl, Padova, Italy
[2]CRIBI, University of Padova, Padova, Italy

Motivation

Searching similarities in protein sequence databases has become a routine procedure in molecular biology. The algorithm of Smith-Waterman has been available for more than 25 years; it is based on dynamic programming and performs an exhaustive search, being able to identify the exact best alignment of each database sequence against a query sequence. Since protein and DNA databases have become considerably large, heuristic approaches, such as those implemented in FASTA and BLAST, tend to be preferred for searching databases. Thus, the loss of accuracy is balanced by a faster execution time. The main motivation of our work is the development of new software/ hardware solutions to implement the Smith-Waterman algorithm on personal computer at a low cost and high performance.

Methods

Here, we describe a new approach to bio-sequence database scanning using Nvidia CUDA compatible graphics card (GeForce 8800) as an efficient hardware accelerator of the Smith-Waterman algorithm. Unlike previous approaches based on OpenGL, this is the first implementation directly running on the GPU hardware without any conversion of the problem to the graphical domain. But CUDA environment imposes specific development and memory models which require state of the art strategy in designing the implementation. When programmed through CUDA, the GPU is viewed as a highly multithreaded co-processor. Its threads are grouped in warps, which compose blocks.

Finally a grid of blocks must be managed by the software developer. Furthermore the memory model of the CUDA device is a composition of six different kinds of memory. So the proper design of the memory usage by the application can dramatically impact the performance of the solution.

We explored different software architectures to determine which will be the optimal in terms of the above restrictions. This way the implementation of the algorithm extracts the maximum performance of a single GeForce 8800 GTX board and allows to obtain a near to linear speed-up as additional boards are added. It makes possible low-cost GPUs to be applied as accelerators having performance over the GCUPS range, never reached up to now on commodity hardware platforms.

We also discuss the potential approaches to improve the performance of heuristic methods like BLAST using CUDA compatible graphics hardware as efficient accelerators.

Results

Comparative tests have been done considering: our GPU implementation using a single GPU, a CPU Smith-Waterman implementation, FASTA and BLAST.

The hardware used is composed by an Nvidia GTX 8800 and a Pentium IV 3.0GHz. The tests consist in alignment of seven different sequences (lengths of 63, 127, 255, 361, 511, 1023, 2047 amynoacids) with the Swiss-prot database (December 2006 - 250296 sequences).

Our solution achieves a speed-up of more than 30 times over a straight forward CPU implementation. Furthermore it is about 3 times faster than previous GPU approaches in the most useful range of the sequence lengths. FASTA is from 9 to 14 times slower than our implementation while we are able to achieve performances comparable with BLAST although this algorithm, like FASTA, benefits of a faster, but less accurate, heuristic approach.

Finally our approach allows for a linear speed-up as additional GPUs are added.

Image: http://www.finorient.com/pics/CudaPerform.jpg

Email: smanavski@elaide.com