GEAgpu: Improved alignment of spliced DNA sequences to genomic data using Graphics Processing Units

Svetlin A. Manavski, Alessandro Albiero, Claudio Forcato, Nicola Vitulo, Giorgio Valle

CRIBI, University of Padova, Padova, Italy E-mail: svetlin.manavski@cribi.unipd.it

http://bioinformatics.cribi.unipd.it/cuda

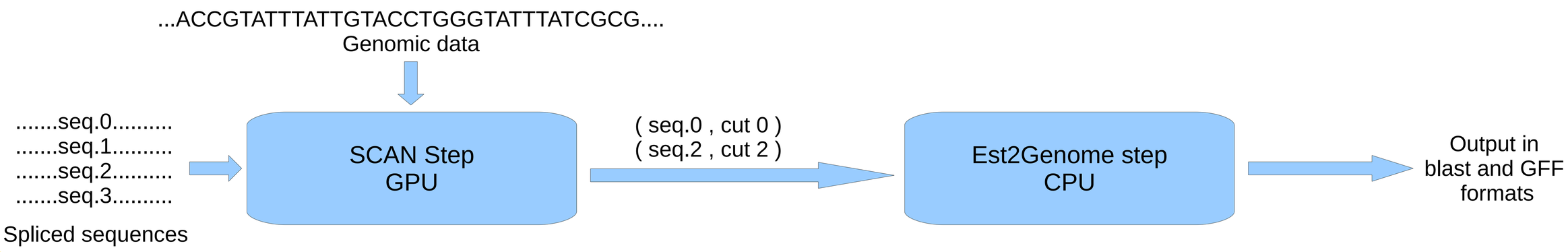
Introduction

The recent DNA sequencing technologies like Illumina-Solexa, Solid, etc. are able to produce tens of millions of sequences in a single run. The research community already produced some tools like Maq and Soap[2] to address the problem of aligning these large sets of short sequences on the genome. One of the most critical steps is the alignment of splice sites both in terms of sensitivity and time requirement. GEAgpu combines the power of graphics processors with a specifically designed algorithm for aligning splicing sites, thus providing a very efficient tool that to our knowledge outperforms any other software available for spliced sequence alignment.

GEAgpu Algorithm

GEAgpu (Genome to Est Alignment on Graphics Processing Units) is a pipelined composition of two main steps:

1. SCAN: selection of all the candidate-areas of the genomic sequence which may potentially align to the query sequence
2. Est2Genome: a Smith-Waterman extension for local alignment of two sequences considering exons and introns [1]

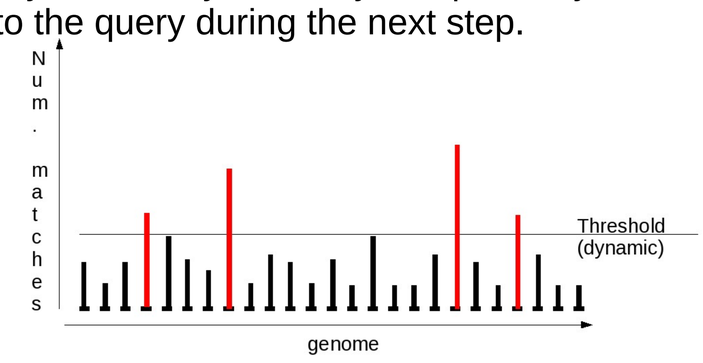


The SCAN step

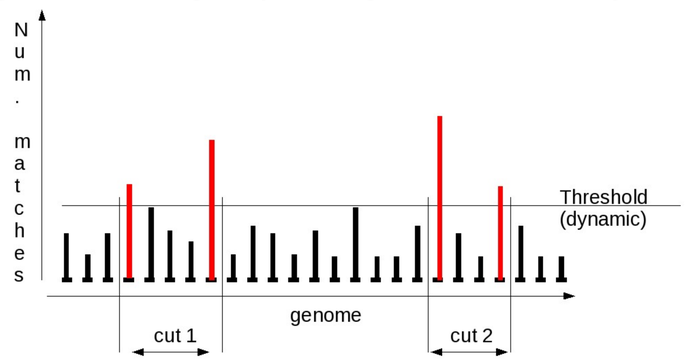
The goal of the SCAN step is to identify all the areas of the genome which may produce a potentially good score in the alignment to the query sequence. Its trade-off is between selecting larger areas to guarantee optimal results and cutting out as much data as possible to improve the speed of the following Est2Genome step.
The SCAN works on indexed genome for maximum performance. The index is produced off-line one time only and saved to the disk as a table of all the genomic positions of all the possible words. To keep the sensitivity as high as possible, GEAgpu uses very short word sizes (4 or 6 bases), which are much shorter than most of the other available alignment tools.

The SCAN overlaps the query to every position of the genomic sequence. For each position it computes the number of matching words and considers it to be the score of the position. Then it ignores all the areas of the genome having scores lower than a pre-defined threshold. The threshold may be also dynamically computed by the software. Finally, the resulting areas are grouped together to identify cuts of the genome to be aligned to the query during the next step.

1)



2)



CUDA and Nvidia G80 [3]

The SCAN step of GEAgpu is implemented in the CUDA programming environment by NVidia. CUDA stands for Compute Unified Device Architecture. It is a hardware and software architecture for issuing and managing computations on the GPU. CUDA is supported by every GPU of the G80 generation and later ones.
From the developer's point of view CUDA is designed as an extension of the C/C++ programming language. CUDA features a parallel data cache or on-chip shared memory with very fast general read and write access. The GeForce 8800 GT graphics card, used in this project, has a total of 112 processing units running on 1,5 GHz and 512 MB of on board DRAM.

Performance tests

The performance was tested by comparing the results to the following packages: Blast, Blat and Soap.The testing set was composed by 50 000 unique sequences generated by Illumina-Solexa sequencer (33 bp each). These were aligned to *Vitis vinifera* genome which size is of 510 Mbp. Although all the 3 packages are useful when aligning sequences, only Blat could be directly compared to GEAgpu because Blast and Soap are not designed to consider splice sites. All the packages were run with a word size of 6 and the results were filtered to keep only alignments which contain at least 30 identities per query. The hardware platform was a Pentium Q6600 2,4 GHz, RAM 2 GB, 2 x nVidia GeForce 8800 GT.

	Runtime (min)	# of alignments	# of queries aligned	Runtime factor
GEAGpu	8	87,596	46,382	1
Blast	2,297	56,799	39,073	277
Blat	388	75,820	43,210	47
Soap	18	69,424	43,513	2

References:
(1) Mott, R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. 13: 477-478
(2) Ruiqiang Li, et. al. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008 24: 713-714
(3) http://www.nvidia.com/object/cuda_home.html