BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 22 2014, pages 3272–3273
doi:10.1093/bioinformatics/btu535

Sequence analysis

Advance Access publication August 13, 2014

gCUP: rapid GPU-based HIV-1 co-receptor usage prediction for next-generation sequencing

Michael Olejnik[1], Michel Steuwer[1], Sergei Gorlatch[1] and Dominik Heider[2,*]

[1]Institute of Computer Science, University of Muenster, 48149 Muenster and [2]Department of Bioinformatics, University of Applied Sciences Weihenstephan-Triesdorf, 94315 Straubing, Germany

Associate Editor: Inanc Birol

ABSTRACT

Summary: Next-generation sequencing (NGS) has a large potential in HIV diagnostics, and genotypic prediction models have been developed and successfully tested in the recent years. However, albeit being highly accurate, these computational models lack computational efficiency to reach their full potential.

In this study, we demonstrate the use of graphics processing units (GPUs) in combination with a computational prediction model for HIV tropism. Our new model named gCUP, parallelized and optimized for GPU, is highly accurate and can classify >175 000 sequences per second on an NVIDIA GeForce GTX 460. The computational efficiency of our new model is the next step to enable NGS technologies to reach clinical significance in HIV diagnostics. Moreover, our approach is not limited to HIV tropism prediction, but can also be easily adapted to other settings, e.g. drug resistance prediction.

Availability and implementation: The source code can be downloaded at http://www.heiderlab.de

Contact: d.heider@wz-straubing.de

Received on May 2, 2014; revised on July 18, 2014; accepted on August 1, 2014

1 INTRODUCTION

Next-generation sequencing (NGS) technologies are currently moving into the field of clinical diagnostics, as such enabling the clinicians to analyze not only the majority variants in HIV quasispecies but also smaller fractions and minority variants (Archer *et al*., 2012). This is especially important for HIV drug resistance prediction, as small fractions of resistant variants can rise during antiviral treatment and can lead to failure of antiviral therapy. An example for such a therapy failure due to minority variants can be found in Dybowski *et al*. (2010b). We analyzed NGS data from four different patients during treatment with an CCR5-antagonist. These drugs bind specifically to the CCR5 co-receptor on the host cells and, by this, prevent viral entry. However, there exist two classes of HIV viruses, the CCR5-using viruses and the CXCR4-using viruses. Obviously, these CCR5-antagonists have no effect on the latter ones. The binding of HIV to the co-receptor is mainly mediated by the so-called V3 loop (Hwang *et al*., 1991), which is part of the surface protein gp120 of HIV. There exist some computational models for HIV tropism prediction, e.g. Dybowski *et al*. (2010a) and geno2pheno (Lengauer *et al*., 2007). These models make their predictions based on genotypic information of HIV, namely the sequence of the V3 loop. Albeit these computational models have been demonstrated to give reliable results on NGS, they all lack computational efficiency with regard to computing time, which is the main bottleneck to clinical relevance of these technologies. Next-generation sequencing technologies can generate up to billions of reads for a given sample. For instance, with 454 technology one can generate around 1 million reads, while with the use of the Illumina technology one can generate up to 3 billion reads (Liu *et al*., 2012).

With this large number of reads, minority variants can be reliably detected. However, the huge amount of data generated needs to be efficiently analyzed in a reasonable period. For instance, T-CUP, albeit highly accurate, needs around 50 days on a single CPU for 40 million reads from a 454/Roche GS FLX sequencing run.

In this study, we introduce and benchmark our new computional prediction model named gCUP. This model is based on our recently developed method T-CUP (Dybowski *et al*., 2010a; Heider *et al*., 2014), but was redeveloped, parallelized and optimized for the use on graphics processing units (GPUs). In Heider *et al*. (2014), we demonstrate the accuracy of our T-CUP model for the prediction of HIV-1 tropism compared with other available models. On an independent dataset, T-CUP outperformed Geno2Pheno as well as the model of Bozek *et al*. (2013). gCUP and T-CUP give identical predictions and thus the accuracy of the model is not compromised by using GPUs. By harvesting the power of GPUs and optimizing the use of their fast local memory, gCUP can drastically reduce the runtime and process the same 40 million reads in just 4 min using one modern GPU.

2 IMPLEMENTATION

For our GPU-based implementation of gCUP, we used OpenCL on an NVIDIA GeForce GTX 460 with 336 CUDA cores and peak performance of 907 GFLOPS (Giga Floating Point Operations Per Second). gCUP is an R package and uses the C interface of R to manage the computation on the GPU. It seamlessly transfers the data to and from the GPU.

gCUP uses as input DNA or amino acid sequences in `FASTA` format. The results are given as pseudo-probabilites that a given sequence belongs to a CXCR4-using virus and reported in a csv file. The following steps are performed to predict the HIV-1 co-receptor usage for a—possibly large—set of input sequences:

- Preparation and quality control (QC) of the input sequences and translation into proteins is performed on the CPU.

*To whom correspondence should be addressed.

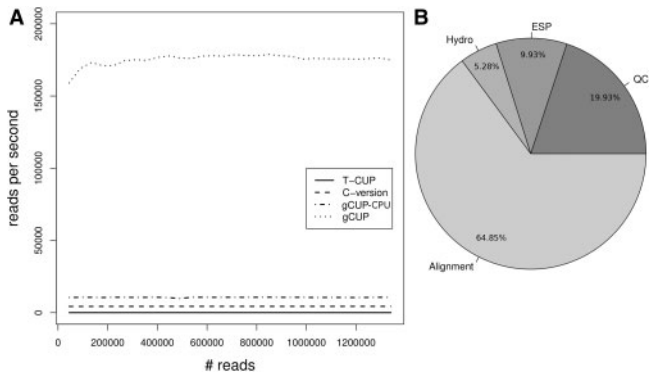

Fig. 1. (**A**) Performance comparison for different input data sizes. On the *x*-axis, the number of reads for a given dataset and on the *y*-axis the number of reads per second that are predicted are shown. (**B**) Percentage of the runtime for the different processing steps within gCUP is shown. QC: quality control; Hydro/ESP: classification with the hydrophobicity and ESP classifier (Heider *et al.*, 2014); Alignment: V3 alignment to the reference sequence and extraction

Some parts of the preparation are implemented in C for performance reasons.

- Alignment of the sequence against reference V3 sequence and V3 extraction are using the Gotoh algorithm (Gotoh, 1982). Our implementation is inspired by Ligowski and Rudnicki (2009) and optimizes the use of the fast but small local GPU memory for achieving high performance.
- Encoding into hydrophobicity and electrostatic potential (ESP) values and interpolation (Dybowski *et al.*, 2010a) as well as the classification is done entirely on the GPU. We have built an efficient GPU implementation of the Random Forest that avoids repetitive accesses to the slow GPU global memory and is streamlined to avoid divergent control flow of threads.

3 RESULTS AND DISCUSSION

Besides the comparison between T-CUP (written in R) and gCUP (written in OpenCL), we also developed a C-based version of T-CUP (in the following referred to as *C-version*) to suppress the negative performance effect of R. Additionally, we also tested gCUP on an Intel Core i5-3550 (referred to as *gCUP-CPU*), as the OpenCL implementation is not restricted to the use with GPUs. The Intel Core i5-3550 has four cores with each being clocked at 3.34 GHz.

As shown in Figure 1A and Table 1, gCUP outperforms all the other implementations of T-CUP with regard to number of sequences per second. While T-CUP is able to predict only around 9 sequences per second, gCUP is able to classify >175 000 on average per second. This improvement is neither based only on the fact that gCUP is written in C, as can be seen by comparison between gCUP and the C-version of T-CUP (175 245 versus 4333 sequences per second), nor it is based solely on the OpenCL implementation (10 541 versus 175 245 sequences per second). The high performance increase is based almost completely on the massive parallelization and optimization targeting GPU architecture. We also analyzed the runtime with regard to

Table 1. Comparison between T-CUP, C-version, gCUP and gCUP-CPU

Methods	T-CUP	C-version	gCUP-CPU	gCUP
Mean reads per second	9	4333	10 541	175 245
Variation coefficient	–	0.004	0.015	0.022

Note. The mean number of reads per second as well as the variation coefficient of several runs with different sizes (40 000 to 1 300 000 reads) are shown

the different steps that are performed within gCUP, namely QC, alignment of the reads against reference V3 sequence and V3 extraction and hydrophobicity and ESP classification (Fig. 1B). Most of the runtime is needed for the alignment and extraction of the V3 loop from the reads (64.85%). QC makes up the second largest part with 19.93%, while hyrophobicity and ESP classification account only 5.28 and 9.93%, respectively. To the best of our knowledge, this study demonstrates for the first time the successful use of GPUs for HIV-1 tropism prediction based on NGS data. For our solution, no expensive computing clusters are needed. Instead, we used a typical desktop PC with an Intel Core i5-3550 and an NVIDIA GeForce GTX 460, which makes our solution attractive for a broad user community. Moreover, our solution is not limited to tropism prediction. It can also be used for other prediction models, e.g. protease inhibitor resistance and reverse transcriptase inhibitors with comparable speed-ups (Heider *et al.*, 2013; Lengauer and Sing, 2006), as the current implementation can handle sequences up to a maximum length of 65 535 amino acids, which is enough for all known proteins.

Funding: This work was supported by the Straubing Center of Science.

Conflict of interest: none declared.

REFERENCES

Archer,J. *et al.* (2012) Use of four next-generation sequencing platforms to determine HIV-1 coreceptor tropism. *PLoS One*, **7**, e49602.

Bozek,K. *et al.* (2013) Analysis of physicochemical and structural properties determining HIV-1 coreceptor usage. *PLoS Comput. Biol.*, **9**, e1002977.

Dybowski,J.N. *et al.* (2010a) Prediction of co-receptor usage of HIV-1 from genotype. *PLoS Comput. Biol.*, **6**, e1000743.

Dybowski,J.N. *et al.* (2010b) Structure of HIV-1 quasi-species as early indicator for switches of co-receptor tropism. *AIDS Res. Ther.*, **7**, 41.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Heider,D. *et al.* (2013) Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*, **29**, 1946–1952.

Heider,D. *et al.* (2014) A simple structure-based model for the prediction of HIV-1 co-receptor tropism. *BioData Mining*, **7**, 14.

Hwang,S.S. *et al.* (1991) Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science*, **253**, 71–74.

Lengauer,T. and Sing,T. (2006) Bioinformatics-assisted anti-HIV therapy. *Nat. Rev. Microbiol.*, **4**, 790–797.

Lengauer,T. *et al.* (2007) Bioinformatics prediction of HIV coreceptor usage. *Nat. Biotechnol.*, **25**, 1407–1410.

Ligowski,L. and Rudnicki,W. (2009) An efficient implementation of Smith Waterman algorithm on GPU using CUDA, for massively parallel scanning of sequence databases. In: *Proceedings of the 2009 IEEE International Symposium on Parallel & Distributed Processing*, Rome, Italy, pp. 1–8.

Liu,L. *et al.* (2012) Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, **2012**, 251364.