Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 596–605

Procedia
Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Fast Genome-Wide Third-order SNP Interaction Tests with Information Gain on a Low-cost Heterogeneous Parallel FPGA-GPU Computing Architecture

Lars Wienbrandt*, Jan Christian Kässens*, Matthias Hübenthal, and David Ellinghaus

Institute of Clinical Molecular Biology, University Medical Center Schleswig-Holstein, Campus Kiel, Kiel University, Germany
{l.wienbrandt,j.kaessens,m.huebenthal,d.ellinghaus}@ikmb.uni-kiel.de

Abstract

Complex diseases may result from many genetic variants interacting with each other. For this reason, genome-wide interaction studies (GWIS) are currently performed to detect pairwise SNP interactions. While the computations required here can be completed within reasonable time, it has been inconvenient yet to detect third-order SNP interactions for large-scale datasets due to the cubic complexity of the problem.

In this paper we introduce a feasible method for third-order GWIS analysis of genotyping data on a low-cost heterogeneous computing system that combines a Virtex-7 FPGA and a GeForce GTX 780 Ti GPU, with speedups between 70 and 90 against a CPU-only approach and a speedup of approx. 5 against a GPU-only approach. To estimate effect sizes of third-order interactions we employed information gain (IG), a measure that has been applied on a genome-wide scale only for pairwise interactions in the literature yet.

Keywords: GWIS, epistasis, three-way SNP interactions, information gain, mutual information, entropy, information theory, hybrid computing, heterogeneous architectures, FPGA, GPU

1 Introduction

In the last ten years genetic research became dominated by genome-wide association studies (GWAS). GWAS were successful in revealing hundreds of SNP-phenotype associations for complex disease traits [18]. High-throughput genotyping methods allow reliable genotyping of millions of single-nucleotide polymorphisms (SNPs) for statistical association testing with a disease in thousands of individuals. In 2007, a first flagship GWAS has been published by the Wellcome Trust Case Control Consortium comprising approx. 3,000 healthy controls and 14,000 cases evenly distributed over seven diseases genotyped for approx. 500,000 SNP markers [22].

[*] corresponding authors, shared first authorship

10.1016/j.procs.2017.05.210

Recent findings reveal that besides simple SNP-phenotype associations, the interaction of genetic markers may also play a significant role in the etiology of complex diseases [6], introducing the research field of genome-wide interaction studies (GWIS). However, an exhaustive computation of interactions on genome-wide datasets is computationally demanding, as for their detection a test statistic has to be calculated for each possible SNP combination. Still, several methods that exhaustively search for pairwise interactions exist, including BOOST [24] and MB-MDR [2, 23]. Since the absolute runtime is at the edge of reasonableness many tools harness accelerator architectures such as GPUs (e.g. GBOOST [26] or GWIS [8]) or perform a heuristic search by eliminating unlikely SNPs in advance (e.g. by clustering approaches [17] or machine learning techniques [25]).

Currently, genetic research is breaking new grounds by exploring the domain of third-order interactions. Due to cubic problem complexity the runtime dramatically increases for genome-wide third-order interaction tests, making standard analyses impractical. Nevertheless, third-order interactions have been proven to play a significant role in the development of complex diseases, e.g. for tuberculosis [3], and hundreds of further available GWAS datasets await exploration of higher-order interactions. Besides heuristic approaches [9, 17], exhaustive methods testing all possible SNP triples have been proposed (e.g. CPM [19], RPM [4] and MB-MDR [2, 23]) but are inconvenient for larger datasets. Accelerator architectures combined with basic information theoretic measurements such as mutual information on GPUs [7] or FPGAs [14] introduce reasonable runtimes with the disadvantage of a less powerful statistic.

In this paper, we introduce a novel approach for accelerating and improving exhaustive third-order interaction detection on a genome-wide scale, firstly, by utilizing a heterogeneous architecture composed of a GPU and an FPGA accelerator included in a desktop PC, and secondly, by implementing the more powerful information gain measurement. Heterogeneous platforms using GPU and FPGA technology have recently been positively evaluated in cryptanalysis [15], astrophysics [16], and image processing [21]. We have also presented a similar platform in [13], but for pairwise interaction detection and with different problem partitioning especially regarding result filtering. Here, we demonstrate that our FPGA-GPU hybrid system for third-order interaction detection outperforms the GPU-only approach GPU3SNP [7] by a speedup factor of 5, and a CPU-only approach by a factor of 90. Furthermore, by using only low-cost off-the-shelf components, we achieve a superior performance-to-cost ratio.

2 Third-order SNP Interaction Measurement

2.1 Contingency Tables

In order to perform an entropy-based statistical test on a 3-tuple of genetic markers, the main task is the creation of so-called contingency tables. We focus on typical GWAS datasets that consist of two groups of samples, namely cases (affected by the disease of interest) and controls (unaffected by the disease of interest). Both groups are genotyped at a set of marker positions that usually carry known SNPs or other genetic variants. Here we consider biallelic markers, which is a common use case, i.e. a genotype may appear as homozygous wild (w), heterozygous (h) or homozygous variant (v) type.

For each 3-tuple of genotyped genetic markers in the collected dataset (further simply referred to as SNP triple) a contingency table is created for each disease group, i.e. one for cases ($l = 1$) and one for controls ($l = 0$). Each table contains the number of samples that share a common property of genotypes at the marker positions. Thus, a contingency table contains $3 \times 3 \times 3 = 27$ counters, one for each possible combination of genotypes (see Fig. 1).