Dr Adam Kowalczyk

Principal Scientist
NICTA

adam.kowalczyk@nicta.com.au

Dr Adam Kowalczyk is currently a principal researcher in Victorian Research Laboratories of National ICT Australia (NICTA). He leads projects in molecular medicine and biology, leveraging many years of research and commercial experience in pure and applied mathematics, mathematical physics, artificial intelligence, telecommunications and, recently, bioinformatics.

GPU enables search for 2-way and 3-way interactions in GWAS

Adam Kowalczyk[a,b], Qiao Wang[a,b], Fan Shi[a,b], Andrew Kowalczyk[a], David Rawlinson[a], Benjamin Goudey[a,b], Richard Campbell[a], Herman Ferra[a]

[a]NICTA, Victorian Research Laboratories, The University of Melbourne, Parkville, VIC 3010, Australia
[b]Computing and Information Systems, The University of Melbourne, Parkville, VIC 3010, Australia

SUMMARY

Genome-wide association studies (GWAS) probe millions of DNA loci in an attempt to associate DNA mutations with a given disease. Complex aetiologies of many common diseases involve combinations of different genes which require individual evaluation of trillions (non-additive) combinations of loci for association in an average size study. We have developed solutions using a single GPU to evaluate association of each and every one bivariate feature within minutes (available via free webserver). Although an exhaustive tri-variate analysis requires currently a medium size GPU cluster, many focused tri-variate analysis tasks can be accomplished routinely on a single GPU within hours of computation.

INTRODUCTION

In recent years, GWAS has been considered a fundamental instrument in unveiling the genetic aetiology of non-Mendelian complex diseases. More than 600 GWAS have been conducted for 150 different diseases and traits. Current genotyping technologies such as Affymetrix 6.0 allow GWAS assays of several million single-nucleotide polymorphisms (SNPs). In order to reveal the underlying biological mechanisms of disease, most analytical methods analyse each SNP individually for association with disease[1]. However, interactions between loci are believed to contribute to complex diseases with non-negligible joint effects[2], even while each SNP may show little effect independently[3]. As recently as 2010 it was considered technically impractical to exhaustively search for second-order (bivariate) interactions without access to state of the art computing facilities due to the large search space (of ~1011 for all pairs and ~1016 for all triples in typical GWAS with 500,000 SNPs). In response, researchers proposed several pre-filtering and stochastic partial search methods using univariate analysis for cutting the number of probes to be considered for multi-locus investigation. However, the worry remains that such methods may miss critical genuine interactions which may show only very weak marginal effects[3,1,5].

DESCRIPTION

Spectacular progress in commodity computing technology in the last few years has led to development of a number of algorithms capable of exhaustive bivariate analysis not only on moderate computer clusters but also on standard desktop computers equipped with General Purpose Graphics Processing Units (GPUs). In order to fully exploit the potential of these devices for medical and biotechnological research there is a need for efficient software tools that reduce implementation and performance difficulties, so researchers can focus on comparison and evaluation of results rather than software tools development and low level algorithm tweaking; see a recent elaboration of this point in[6]. It has been demonstrated[7] that the number of novel putative epistatic loci can be detected using such techniques.

In this talk we present an efficient library of GPU kernels that can be used for fast implementation of any bivariate GWAS statistics that can be derived from contingency table counts. In order to illustrate the point we have implemented nine different algorithms from the literature. All algorithms can execute exhaustive search on typical case/control GWAS (500K SNPs, 5K samples) within 10 minutes. This performance makes comparative analysis of different statistical methods easy. These results are also significantly improved compared to original implementations of the nine algorithms considered. Speedup factors of over 300 are observed compared to some original GPU implementations in literature and even larger factors of over 10,000 are seen with respect to the CPU implementations, e.g. a popular Fast Epistasis algorithm in PLINK 1.07 software package.

Consider that timing scales quadratically with the density of genotyping markers used. Future high-density SNP arrays will include up to 5 million SNPs, and forthcoming GWAS based on NGS data will have even higher marker density and may include other technology such as methylation markers. Together, these expectations mean that the computational burden of exhaustive bivariate analysis will continue to be challenging: between one hundred and one thousand times more complex than existing GWAS. Thus our GPU approach, which

efficiently applies a battery of statistical tests to exhaustive search of all SNP pairs in minutes for current GWAS data, will still require hours or days for near-future data. This is achievable on a single GPU-equipped desktop or laptop computer, but time scales down accordingly if GPU clusters are used. Additionally, the users can define and add their own statistics to our platform and make use of our high performance library that generates contingency tables and ranks scores. To facilitate this we have insured compatibility with popular input data formats, a common interface for defining statistical tests. The runtime of ~10 minutes for a typical dataset and computer is fast enough that researchers can experiment with methods interactively, reviewing the effects of varied algorithms almost immediately.

The practical consequences of such an improvement in productivity are not a matter of degree. By enabling researchers to conduct GWAS experiments in less time than a coffee break it becomes possible to focus effort on statistical methods and results rather than avoiding performance bottlenecks. New ideas can be implemented and evaluated in very fast cycles, without the need to book time on shared high end computing resources.

Most recently, we have extended GWIS to exhaustive search for 3-way interactions, a previously impossible computational task. Using our methods, an exhaustive 3-way analysis of Celiac disease GWAS from UK containing ~310K SNPs and 2200 samples using a cluster of 200 GPUs requires 7 days of computing time. To our knowledge this is the first time such an analysis has been shown to be practical. The runtime reduces significantly for more targeted analysis, for example a specific DNA region or a preselected set of SNPs. Exhaustive filtering through all SNP-triplets in ~2500 SNPs, including the extended MHC region, requires <3 minutes on a standard PC with a single GTX470 NVIDIA GPU.

CONCLUSION

In conclusion, analysis of two-way and three-way interactions in modern GWAS using multiple methods is practical today. Once such analyses are practiced in labs around the world faster progress in unveiling the genetic aetiology of complex diseases may result. To date, it has been difficult to compare methods across a range of datasets due to implementation difficulties and prohibitive runtime. Many existing benchmark studies were forced to use only small size, typically synthetic, datasets. What we claim is a paradigm shift, towards routine usage real life data for methods development, benchmarking and then “production” deployment of novel GWAS data analysis paradigms.

REFERENCES

1. Cordell, 2009, Nat.Rev.Genet, 2009, 10(6), 392–404.
2. Marchini et al., 2005, Nat Genet, 37(4), 413–417.
3. Zhang and Liu, 2007, Nature Genetics, 39(9), 1167–1173.
4. Culverhouse et al., 2002, Am. J. Hum. Genet., 70, 461-471
5. Zhang et al., 2010, J. Comp. Biol. 17(3), 401–415.
6. Wilson et al. (2014, PLoS Biol, 12(1).
7. Godey et al. (2013), BMC Genomics 14.