Master's project proposal: Sparse matrix computations in genome-wide association studies

Although we today have a map of the human genome[1] and can take DNA samples in a simple and efficient way, there are many genetically determined human features and diseases that we cannot identify from the DNA sample. The problem is that most research analyzes single DNA mutations and not combinations.

The newly started company GenoKey[2] has developed a new method for identifying critical genotype *combinations* that may cause genetic disorder. The method has shown interesting results when analyzing a set of around 800 DNA mutations pointed out by doctors to be relevant for a specific disease [Koefoed et al. 2011]. To make the method genome-wide, such that it can generate even more surprising results in the study of genetics, the method must be able to scale such that hundreds of thousands of DNA mutations may be analyzed. We offer a Master's project in cooperation with GenoKey where the challenge is to speed-up the computation methods needed for this hard combinatorial problem. A key element in speeding up the computations is to use the graphics processing unit (GPU). The following is an introduction to the research area and the computational problem.

Project background

The DNA (deoxyribose nucleic acid) sequence[3] which is found in the human genome consists of millions of chemical base pairs[4] (two nucleotides[5] connected via hydrogen bonds). DNA mutations are recorded as SNPs (single-nucleotide polymorphisms[6]) which are either homozygous, heterozygous, or double-heterozygous. We enumerate these three cases by the numbers 0, 1, 2 such that each SNP can be encoded by two bits. A homozygous SNP in a person signals that this person has inherited the normal nucleotide from both parents. A heterozygous SNP signals that one nucleotide deviates from the normal one. A double-heterozygous SNP signals that both nucleotides are deviating. Analysis of SNPs and combinations of

[1] http://en.wikipedia.org/wiki/Human_genome
[2] http://genokey.com/
[3] http://en.wikipedia.org/wiki/DNA_sequence
[4] http://en.wikipedia.org/wiki/Base_pair
[5] http://en.wikipedia.org/wiki/Nucleotide
[6] http://en.wikipedia.org/wiki/Single-nucleotide_polymorphisms

SNPs in patients with polygenic diseases is the key to personalized medical treatment. The combinatorial problem is, however, immense. So the challenge is to do combinatorial data analysis of SNPs in patients as compared to control persons.

The first thing we should be able to do is to count the number of patients versus controls with a particular SNP of a particular value (0, 1, or 2). To handle the large number of persons and SNPs, we employ the graphics card.

In a genome-wide study, we would consider half a million SNPs or more. In such a study, we need a way of counting only a selection of SNPs of a particular value. One way to do this is to have three Boolean matrices with persons along the vertical axis and SNPs along the horizontal axis. The first matrix A_0 would point out homozygous SNPs for all persons, the second A_1 would point out heterozygous SNPs, and the third A_2 would point out double-heterozygous SNPs. These matrices could be stored in a sparse format such as the disjunctive normal form with only indices to the elements that are valid. Another option is one person-SNP matrix with two bits for each element. Assuming that most people have normal SNPs, this matrix would probably be highly suitable for storage in the coordinate list format (COO)[7]. The particular way of doing the selection and counting would depend on the choice of format. Once efficient selection and counting is established, many other core operations in the combinatorial data analysis are likely to benefit tremendously.

Contact

Jeppe Revall Frisvad, Associate Professor, DTU Informatics, jrf@imm.dtu.dk
Peter Falster, Docent Emeritus, DTU Informatics, pfa@imm.dtu.dk
Gert L. Møller, CEO, GenoKey ApS, glm@genokey.com

References

KOEFOED, P., ANDREASSEN, O. A., BENNIKE, B., DAM, H., DJUROVIC, S., HANSEN, T., JORGENSEN, M. B., KESSING, L. V., MELLE, I., MØLLER, G. L., MORS, O., WERGE, T., AND MELLERUP, E. 2011. Combinations of SNPs related to signal transduction in bipolar disorder. *PLoS ONE 6*, 8 (August), e23812.

[7] http://en.wikipedia.org/wiki/Sparse_matrix