WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME

Centrum Studiów Zaawansowanych PW
Center for Advanced Studies WUT

Challenges of Genome Wide Search for Biomarkers

Prof. Adam Kowalczyk

The completion of the first draft of human genome in year 2000 heralded arrival of technologies capable analysing complex organism such as human on fundamental level of DNA. The refinements of last decade made those technologies affordable to the level of practical applications in personalised medicine, animal breeding and agriculture, as the cost of genotyping dropped dramatically. However, this creates new challenges and bottlenecks, namely, the need for techniques capable of data mining of knowledge from such ever expanding wealth of data.

Even for the simplest task of search for multi-loci biomarkers we are forced to consider explicitly billions and trillion dimensions with a very limited utility for local averaging. This requires development of novel efficient techniques for variable filtering founded on solid extensions of statistical techniques well beyond classical boundaries. These searches for biomarkers are so large that they were declared completely impractical and these claims were refuted only recently by practical deployment of recent cluster computing technologies. However, the current challenge is to deploy novel, cost efficient computational platforms as the practical solutions need to be deployed in thousands of laboratories around a globe on a routine bases in order to make a real change. We and others demonstrated feasibility of such solutions recently.

The story is far from over yet. Next come proper biological and medical interpretation of results: linking to external information sources, such as ever expanding genome browsers, pathways and functional genomic databases, etc. This necessitates development and deployment of novel artificial intelligence solutions, e.g. text mining tools capable of scanning literature for relevant connections.

In this series of lectures we outline a cross-section of results developed at Victoria Laboratory of National ICT Australia, in Melbourne, split in four presentations.

Lecture co-financed by the European Union in scope of the European Social Fund

HUMAN CAPITAL
HUMAN – BEST INVESTMENT!

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

1. Challenges of Genome Wide Association Studies

It is estimated that over $1B has been invested globally in development of data for genome wide association study (GWAS) for over 500 diseases. This wealth of data is still waiting for proper analysis and identification of mutations and their combinations critical for disease development or cure form it. In parallel, similar datasets are developed in the context of animal and crop breeding and management. We shall introduce the task and point to literary cosmic-size challenges involved with mining knowledge from such datasets.

2. Development of scanning filters for GWAS

We present challenges and discuss some solution for development of practical scanning filters for searching for interactions in the trillions (10^12) of pairs and over 10^15 triplets of DNA loci in current generation of GWAS data.

3. Paradoxes in learning in high dimensional spaces

We shall discuss a paradox, when computer learns to become a compulsive liar. This happened in practical cancer genomic predictions and can be demonstrated and proved formally. The point is that we need to be very open-minded in processing high dimensional data as our every-day intuitions can be inadequate.

4. Some success stories: personalised tests for cancers

We shall discuss a few practical applications of genomics which made the way from our lab to practical pathology test for cancer as well as tests deployed by some genotyping services used to advice plant breeders and farmers.