Scientists cut through data noise of high-throughput DNA sequencing with mathematical technique

June 17 2013

(Phys.org) —Scientists at A*STAR's Genome Institute of Singapore (GIS) have developed a revolutionary method to quickly cut through noise and generate a unified and simplified analysis of high-throughput biological data from, for example, patient samples. The technique, known as a pre-whitening matched filter, is well known in electrical engineering and widely used in cell phones and radar. This is the first time, however, computational scientists, led by Dr Shyam Prabhakar, Associate Director, Integrated Genomics, GIS, have adapted it to the analysis of high-throughput DNA sequencing data, with surprisingly accurate results. The development was recently published in the prestigious journal, *Nature Biotechnology*.

High-throughput DNA sequencing has revolutionized the study of molecular biology and human disease. The technology has yielded major insights into cancer, infectious diseases, Parkinson's disease and many developmental disorders.

The difficulties facing this technique are the massive amounts of data that are generated. To add to that, it was generally believed that a different method of analysis was required for each type of sequence data . Hence, each new data type was treated as a completely new analysis problem, resulting in a tremendous number of different analytical methods to solve them.

Dr Prabhakar and his team at the GIS, however, discovered that by using the pre-whitening matched filter technique, the results were uniformly better than other existing algorithms at a whole range of analysis tasks. In essence, the technique was applied to accurately detect segments of the genome that stood out from the rest of the sequence data. This was possible because, as lead author Dr Vibhor Kumar quickly realized, the underlying mathematics to the solution of all these analysis problems was the same.

The team was also able to use a variant of the technique to accurately predict gene expression, from epigenomic data. In other words, they could predict the activity levels of genes from data on chemical changes in the genetic material. This is significant especially in clinical settings, since gene expression is difficult to measure by conventional methods in old and degraded tissue samples.

"Our work fits into the pattern of applying engineering solutions to data analytics problems, and we are excited about using our approach to uncover important features of human disease," said Dr Prabhakar. "This discovery will make it a lot easier for scientists to make biological inferences from high-throughput DNA data, particularly in the context of clinical samples from patients."

GIS Executive Director Prof Ng Huck Hui said, ""This is a classic work of high performance computational biology that provides an analytical solution for a complex big data era. With this development, Dr Prabhakar's team brings us one big leap further and faster in scientific high-throughput sequencing work."

Dr Rob Mitra, Alvin Goldfarb Distinguished Professor of Computational Biology and Associate Professor, Department of Genetics at the Washington University School of Medicine said, "This work provides an elegant solution to a ubiquitous problem: separating the signal from the

noise in deep-sequencing datasets. The DFilter algorithm represents a significant advance because it is widely applicable and because it is more accurate than existing algorithms. DFilter can be used to analyze virtually any sequence-tag analysis of DNA binding (e.g. ChIP-Seq, DNASE-Seq, or FAIRE-Seq), and since it uses the mathematically optimal linear discriminant, it was able to outperform all of the existing tools that were developed specifically for each type of assay."

Dr Olli Yli-Harja, Professor in the Department of Signal Processing in Tampere University of Technology, Finland, added, "This is very inspiring for signal processing researchers, as this study demonstrates the great benefit of systematic benchmarking in signal estimation and heterogeneous sample analysis for data generated by next-generation sequencing."

More information: Kumar, V. et al. Uniform, optimal signal processing of mapped deep-sequencing data, *Nature Biotechnology*, 16 June 2013.

Provided by Agency for Science, Technology and Research (A*STAR), Singapore