Formal mathematics underpins new approach that standardizes analysis of genome information

September 11 2013

Credit: AI-generated image (disclaimer)

Researchers in Singapore have developed and tested mathematical tools, or algorithms, that are more accurate and robust than those currently used in analyzing high-throughput genetic sequencing data. The algorithms can determine the location and activity of specific nucleic

acid sequences in a broad range of high-throughput techniques that detect gene–protein interactions. The research group, led by Shyam Prabhakar of the A*STAR Genome Institute of Singapore, also showed they could use the algorithms to generate meaningful results from degraded tissue and tissue constructed from several different cell types.

The rapid expansion in the application of high-throughput sequencing was possible because of a parallel growth in bioinformatics techniques to analyze the huge amount of data that the technique generated. High-throughput sequencing began as a technology for rapidly sequencing whole genomes; now it can detect gene activity, DNA methylation, microRNA binding and interactions between genes, transcription factors and regulatory elements. Each of these different sequencing techniques spawned its own specialized analytical methods, many of them based on heuristics—practical strategies that work but may require optimization.

Prabhakar and his colleagues recognized that almost all sequencing analyses are concerned with solving two major classes of problems, long studied in the fields of signal processing—signal detection and signal strength estimation. Standard mathematical techniques already existed for solving such problems. The researchers therefore adapted these techniques to sequencing analyses. They reasoned that the formal mathematical basis underlying the techniques would allow them to be optimized or tuned. They also realized that the same approaches could be used across a broad range of applications, thus enabling data integration.

The researchers developed two algorithms: DFilter for detecting and locating the binding of regulatory proteins to the genome; and EFilter for estimating gene activity through levels of messenger RNA, the genetic material used as a template for building proteins. Across several sequencing technologies, the researchers benchmarked both algorithms against existing analytical methods. They found that DFilter and EFilter outperformed the more specialized algorithms. The new algorithms also

facilitated the analysis and comparison of multiple and diverse data sets.

Prabhakar and co-workers also used their new algorithms to analyze data from complex, heterogeneous tissue in the embryonic mouse forebrain. They searched for functioning transcription factors and gained useful insights, despite the fact that individual transcription factors could not be assigned to specific cell types.

"We intend to make DFilter and EFilter widely available," says Prabhakar, "perhaps via cloud genomics providers, if all goes according to plan."

More information: Kumar, V., et al. Uniform, optimal signal processing of mapped deep-sequencing data, *Nature Biotechnology* 31, 615–622 (2013). www.nature.com/nbt/journal/v31 … 7/full/nbt.2596.html

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