DETECTING GENOMIC ELEMENTS OF EXTREME CONSERVATION IN HIGHER EUKARYOTES BY INTEGRATION OF HASH MAPPING AND CACHE-OBLIVIOUS IN-MEMORY COMPUTING

Andi Dhroso
Dr. Dmitry Korkin, Thesis Supervisor

Abstract

Genomics is one of the first life science disciplines to enter the era of Big Data, facing challenges in all three dimensions—volume, variety, and velocity. Yet, in spite of a plethora of sequencing data, we are still far from creating a complete encyclopedia of functional and structural elements of the genome. In 2004, an example of this knowledge gap came about when Bejerano and Haussler discovered 481 DNA elements in the syntenic positions of human, mouse and rat genomes that were 100% identical, called the ultra-conserved elements (UCEs).

Our ultimate goal is to provide a comprehensive atlas of the regions of extreme conservation in higher eukaryotes providing insights into the structural organization, function and evolution of these elements. Here, we present a new hybrid approach that integrates the ideas of hash mapping and cache-oblivious in-memory computing. Our algorithm leverages the concept of help-me-help-you, where the data structures are tailored to maximize cache-hit while minimizing cache-miss. As a result, our hybrid algorithm is approximately 800 times faster than the current state-of-the-art method and is scalable to deal with the unassembled genomes. The new hybrid approach has been applied to detect the earliest evidence of extreme conservation by including into the large-scale analysis recently sequenced genomes of coelacanth and lamprey. The integration of efficient software with hardware-optimized approaches has shown to be a promising direction in comparative genomics, allowing scientists to provide even deeper insights into the function and evolution of eukaryotic genomes.