Evidence of the ancient extreme conservation of genomic regions in animal genomes

Andi Dhroso[1], Nathan Johnson[1], Saptaparni Bandyopadhyay[1], Dmitry Korkin[1,2]
Bioinformatics and Computational Biology Program[1], Department of Computer Science[2]
Worcester Polytechnic Institute, Worcester, MA, USA

We have recently witnessed the tremendous progress in evolutionary and regulatory genomics of eukaryotes fueled by hundreds of recently sequenced pant and animal genomes. Yet, we are still far from creating a complete encyclopedia of functional and structural elements of the eukaryotic genome. In 2004, an example of this knowledge gap came about when two groups independently made an intriguing discovery, finding genomic elements that were slowed down through the course of evolution to their extremes. The original elements of extreme conservations included genomic sequences shared between human, rat, and mouse genomes that were fully or nearly identical, allowing at most a few base substitutions, insertions or deletions. In particular, Bejerano and Haussler found 481 DNA sequences that were 100% identical and located in syntenic positions across the three genomes; the sequences were called ultraconserved elements (UCEs).

In spite of the tremendous interest to the subject, little is known about the functions of UCEs, with many UCEs detected in non-coding regions. Some of the original UCEs were found in long non-coding RNAs (lncRNAs) and transposons. Similarly, the origins of ultraconservation are yet to be ascertained. For instance, until recent the existence of such extreme regions beyond tetrapods was not known. One of the main bottlenecks was dependence of the search algorithm on a whole-genome alignment, which was not feasible for a diverse set of genomes or genomes that underwent drastic evolutionary changes, such as in plant species. Our recently developed alignment-free information-retrieval method could find a comprehensive set of extremely conserved elements across multiple genomes, even if they were not in syntenic regions or they were the tandem repeats abundant through genomes. Due to the fact that those elements were defined more broadly than UCEs, we called them Long Identical Multispecies Elements (LIMEs). Using our approach, we have identified and studied the comprehensive sets of LIMEs across six animal species and, for the first time, across six plant species. However, even with this advancement, the question about the origins and evolutionary mechanisms behind the extreme genomic conservation remained open. The primary reason was, again, the computational limits of our approach, in spite of the fact that it leveraged one of the algorithmically fastest search methods, the hash mapping search. The most critical limitations were the inability to process the partially assembled genomes and to search for shorter extreme regions.

We have recently developed a principally new approach to this problem. The idea of this approach is to develop a method that will be optimal from the algorithmic point of view as well as from the point of view of the computational hardware it is run on. Called a cache-oblivious approach, our method is designed to improve a standard hash mapping approach by optimizing the data exchange with the CPU's cache memory. Compared to our original hash-mapping approach, the new cache-oblivious method achieves a remarkable speedup of up to 2,000 times, eliminating all previous computational bottlenecks. Most importantly, by applying our new method to a set of animal genomes including human, mouse, coelacanth, pufferfish, and the recently sequenced but not fully assembled genomes of the elephant shark and lamprey, for the first time we discovered a large group of ancient LIMEs that is conserved throughout all these species, and thus beyond tetrapods, suggesting that the origins of extreme conservation could be as early as 700 MYA. The ongoing analysis of the obtained ancient LIMEs found in exons and introns, as well as in many non-coding regulatory regions has already provided us with insights into their potential functions, including their role in alternative splicing and several complex genetic disorders in humans. In conclusion, to the best of our knowledge this the first cache-oblivious approach in computational genomics, suggesting a paradigm shift in bioinformatics algorithms in order to deal with the exponentially growing data volume.