Public Abstract
First Name:Andrew
Middle Name:John
Last Name:Todd
Adviser's First Name:Michela
Adviser's Last Name:Becchi
Co-Adviser's First Name:N/A
Co-Adviser's Last Name:N/A
Graduation Term:FS 2016
Department:Electrical and Computer Engineering
Degree:MS
Title:Parallel Gene Upstream Comparison via Multi-Level Hash Tables on GPU

The region of DNA immediately in front of a gene body, called an upstream region, contains special sequences, called motifs, that control whether or not a gene is expressed. Unfortunately, these sequences are generally unknown and commonly contain slight variations between related species. A motif-finding framework is proposed here that, given a set of gene upstream regions, performs all-to-all pairwise comparison and identifies all sequences of length k (k-mers) that are common to any pair of upstream regions or differ in at most d characters. The multi-level hash table used optimizes table comparison (rather than hash table insertion or lookup), is highly parallelizable and easily maps onto GPU. Four GPU kernels are proposed to handle theses hash tables, each leveraging a distinct parallelization approach. Moreover, a study of different factors that affect the performance of each is included (the hash function, the number of buckets and the settings of additional implementation-specific parameters). The contribution is merited by experimental results from an average-size yeast genome that show the fastest GPU kernel outperforming an 8-thread, cache-efficient CPU implementation by a factor of approximately 52x.