P77

Genomic-range Overlaps Based on the Maximum Range Length

Ho-Sik Seok 1, Taemin Song 1, Sek Won Kong 2 and Kyu-Baek Hwang 1

1 School of Computer Science and Engineering, Soongsil University, Republic of Korea

2 Informatics Program at Boston Children's Hospital, USA

Abstract

The demand for fast range search is ever increasing with the advancement of high-throughput sequencing technologies. Efficient search algorithms for finding genomic-range overlaps facilitate the annotation of genomic variants in whole-genome sequencing (WGS) as well as detecting overlapping exons with aligned short reads from mRNA sequencing. A majority of fast algorithms for searching the overlaps between a query range (e.g., a genomic variant) and a set of N reference ranges (e.g., exons) has time complexity of $O(k + \log N)$, where k denotes a term related to the length and location of the reference ranges. Here we present a simple but efficient algorithm that reduces k, based on the maximum reference range length. Specifically, for a given query range and the maximum reference range length (e.g., the longest exon length in a chromosome), the proposed method divides the reference range set into three subsets: *always*, *potentially*, and *never overlapping*. Therefore, search effort can be reduced by excluding never overlapping subset. We demonstrate that the running time of the proposed algorithm is proportional to potentially overlapping subset size, that is proportional to the maximum reference range length if all the other conditions are the same. Moreover, an implementation of our algorithm was 13.8 to 30.0% faster than one of the fastest range search methods available when tested on various genomic-range datasets. The proposed algorithm has been incorporated into a disease-linked variant prioritization pipeline for WGS (http://gnome.tchlab.org) and its implementation is available for academic research at http://ml.ssu.ac.kr/gSearch.