ntHash: recursive nucleotide hashing

Hamid Mohamadi, Justin Chu, Benjamin P Vandervalk and Inanc Birol
Canada's Michael Smith Genome Sciences Centre,
British Columbia Cancer Agency, Vancouver, BC, V5Z 4S6, Canada

Background

In bioinformatics, there are many applications that rely on cataloguing or counting DNA/RNA sequences for indexing, querying, and similarity search. These include sequence alignment, genome and transcriptome assembly, RNA-seq expression quantification, and error correction. An efficient way of performing such operations is through the use of hash-based data structures, such as hash tables or Bloom filters. Therefore, improving the performance of hashing algorithms would have a broad impact for a wide range of bioinformatics tools.

Results

Here, we present ntHash, a fast hash method for computing hash values for all possible sub-sequences of length *k* (*k*-mers) in a DNA sequence. The algorithm calculates hash values for consecutive *k*-mers in a given sequence using a recursive approach, in which the hash value of the current *k*-mer is derived from the hash value of the previous *k*-mer. In this work, we have implemented a cyclic polynomial rolling hash function, and adapted it to nucleotide hashing. Particularly, we made use the reduced alphabet of DNA sequences, and handled the reverse complementation efficiently. The proposed method also provides a fast way for calculating multiple hash values for a given *k*-mer without repeating the whole hashing procedure for each value. This functionality would be very useful for certain bioinformatics applications, such as those that utilize the Bloom filter data structure.

Conclusions

Experimental results demonstrate substantial speed improvement over conventional approaches, while retaining near-ideal hash value distribution. Comparison of run time of proposed method with the state-of-the-art general-purpose hash functions demonstrates that ntHash performs over 20x faster than the closest competitor, *cityhash*, the leading algorithm developed by Google.