Wen-Lian Hsu, J Proteomics Bioinform 2018, Volume 11
DOI: 10.4172/0974-276X-C1-111

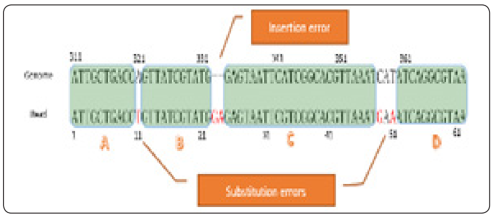
International Conference on

Computational Biology and Bioinformatics

September 05-06, 2018 Tokyo, Japan

Wen-Lian Hsu
Academia Sinica, Taiwan

An ultra-efficient global alignment algorithm for comparing highly similar sequences



Simple pairs and normal pairs. A read sequence can be decomposed into different parts according to the alignment with the genome sequence. A simple pair represents a pair of identical sequence fragments. A normal pair represents a pair of sequence fragments which contains some sequence variations in the alignment.

We present an ultra-efficient global alignment algorithm for comparing similar genomes and for read mapping in Next Generation Sequencing (NGS), which can process long reads as fast as short reads. Furthermore, it can tolerate much higher error rates. Our parallel read mapping algorithm, KART, is 3 to 10 times faster than the well-known Bowtie2 and BWA-MEM algorithm. On pairwise alignment of human genome sequences, the extended KART is 260 times faster than current methods. The same idea has also been applied to RNA-seq, producing DART, a quick and accurate mapping algorithm. (1) KART: a divide-and-conquer algorithm for NGS read alignment (2) DART: a fast and accurate RNA-seq mapper with a partitioning strategy: Besides getting high quality alignment efficiently, our algorithm can simultaneously perform variant calling in about the same amount of time. To achieve the abovementioned objectives, we design a divide-and-conquer alignment strategy giving a query sequence P and one reference sequence Q; identifying all locally maximal exact matches as simple region pairs in sequence Q with sequence P and then clustering the simple region pairs (simple pairs) according to their coordinates in the database to form the bases of global alignment and fixing the overlaps between adjacent simple region pairs and then filling gaps between adjacent simple region pairs by inserting normal region pair (normal pairs) to produce a complete alignment. The crux of the algorithm is that simple pairs can be aligned in linear time and all simple pairs and normal pairs can be aligned independently and in parallel. After dividing the query sequence P sufficiently, those pairs that require gapped alignment only have an average length of 21.

Biography

Wen-Lian Hsu had contributed on the design of graph algorithms and he has applied similar techniques to tackle computational problems in biology and natural language. He has developed a Chinese input software, GOING, which has since revolutionized Chinese input on computer. He is particularly interested in applying natural language processing techniques to understand DNA sequences as well as protein sequences, structures and functions and to biological literature mining. He has designed ultra-efficient alignment algorithms for biological sequences, flexible approximate matching and clustering algorithms for natural language text.

hsu@iis.sinica.edu.tw

Notes: