**SOAP2: an improved ultrafast tool for short read alignment.**

1. Technology Used:

In the paper, they have developed an improved version of SOAP, called SOAP2. The new program uses the Burrows-Wheeler Transformation (BWT) compressed index instead of the seed algorithm that was used in the previous version for indexing the reference sequence in the main memory. The use of BWT substantially improved alignment speed and significantly reduced memory usage.

2. Approach:

Next-generation DNA sequencing technologies, including Illumina/Solexa and AB/SOLiD, have been dominant tools for genomic data collection. Various applications have been developed using these technologies to promote biological research, such as detecting genetic variation through whole genome or target region resequencing, refining gene annotation by whole transcriptome sequencing, profiling mRNA and miRNA expression, and studying DNA methylation. One of the common key data analysis steps of these applications is to align huge amounts of short reads onto a reference genome. New efficient programs have been developed to meet the challenges of such alignment. Among them, SOAP (Short Oligonucleotide Alignment Program) (Li, et al., 2008) has been used widely for these types of analyses due to its fast speed and richness of features.

1. Implementation details:

With further improvement in sequencing throughput and the launch of big research projects, much faster short-read alignment methods are required to handle the data analysis of such large-scale sequence production. For example, the 1000 Genomes project, which aims to create the most detailed and medically useful human is regarded as joint First authors. genetic variation map will generate about 15Tb of the sequence using next-generation sequencing technologies. With even the fastest programs currently available, one would need approximately 1,000 CPU months to align these short reads onto the human reference genome. Additionally, new methods are now needed to support longer reads as the existing methods were primarily designed for very short reads with typical lengths shorter than 50 bp. With improvements in sequencing chemistry and data processing algorithms, the Illumina Genome Analyzer can now generate up to 75~100 bp high-quality reads, and longer reads are expected in the near future. Here we have developed an improved version of SOAP, called SOAP2.

4. Performance:

Suffix trees and suffix arrays are considered the most appropriate methods for indexing DNA sequences, through which only one alignment is needed for repetitive sequences with multiple identical copies in the genome. The complexity in space and time of such index construction has limited such algorithm usage in only small genomes. But the recent development of compressed indexing has reduced the space complexity from O(n) bytes to O(n) bits. Among these is the Burrows-Wheeler Transformation (BWT) (Burrow, 1994), a reversible data compression algorithm, which was found to be the most efficient. The space complexity of BWT is n/4 bytes, and only 1 GB of memory in RAM is required for indexing the whole human genome.

5. Tool name:

SOAP2, SOAP, MAQ, Bowtie.

6. Advantages:

The complexity in space and time of such index construction has limited such algorithm usage in only small genomes. But the recent development of compressed indexing has reduced the space complexity from O(n) bytes to O(n) bits. Among these is the BWT (Burrow and Wheeler, 1994), a reversible data compression algorithm, which was found to be the most efficient. The space complexity of BWT is n/4 bytes, and only 1 GB of memory in RAM is required for indexing the whole human genome.

7. Disadvantages:

Using this in our alignment method, we determined an exact match, by constructing a hash table to accelerate searching for the location of a read in the BWT reference index. For example, if we use a 13mer on the hash, then the reference index would be partitioned into 226 blocks, and very few search interactions are sufficient to identify the exact location inside the block. For inexact (both mismatch and indel) alignment, we applied a ‘split-read strategy’. To allow one mismatch, a read was split into two fragments. The mismatch can exist in, at most, one of the two fragments at the same time. Likewise, we split a read into three fragments to search for hits that allow two mismatches. This enumeration algorithm was used to identify mutation sites on the reads.

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |