Comparing and aligning huge bio-sequence datasets

フリス マーティン[*1)]

1. 目的・背景

Biological sequences (e.g. genomes) contain the information for making living things. For many decades, the main way of analyzing biological sequences has been by comparing and aligning them. This remains true today. Modern tasks include: comparing whole genomes; aligning bisulfite-converted DNA reads to a genome; aligning long, high-error sequences from single molecule sequencers; aligning ancient or degraded DNA; comparing metagenomic DNA to a protein database.

Over the decades, statistically powerful alignment techniques have been developed, including: log likelihood ratio scoring matrices, pair hidden Markov models, and probabilistic alignment. Unfortunately, these methods are rarely used with modern deep sequencing data, perhaps because they are thought to be too slow.

This presentation will describe a software package, LAST, that can do a wide variety of sequence comparison tasks with both powerful statistics and high speed.

2. 研究内容

LAST rapidly finds and aligns similar regions between sequences (図 1). It uses a seed-and-extend approach like BLAST, but gains speed by using *rare* seeds. By using a statistical model of sequence divergence, it calculates the confidence that each column is correctly aligned (lighter shade = low confidence, darker shade = high confidence).

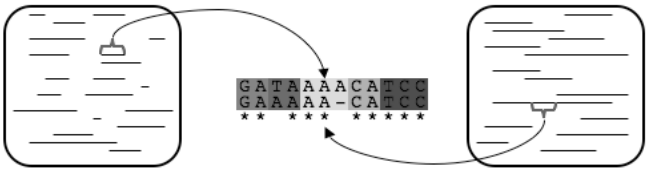

図 1.　Illustration of LAST's function

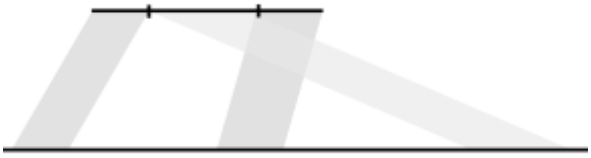

図 2.　Split alignment

LAST can include sequence quality data in its statistics, for more accurate results. It can find weak or strong similarities, between short or long sequences. It can use statistical models for biased sequences (e.g. AT-rich genomes, bisulfite-conversion). It can also align DNA to proteins, allowing frameshifts in the middle of the alignment. An interesting feature is *split alignment* (図 2), where different parts of one query can match disjoint regions of the genome. This is useful for: DNA rearrangements in cancer, spliced RNA (including trans-splicing), and whole-genome comparison.

3. 今後の展開

These techniques provide a general approach to doing many kinds of large-scale sequence comparison. They are broadly applicable to analyzing and answering questions about biological sequences.

*1)独立行政法人産業技術総合研究所