An Approximate Bayesian Approach to Mapping Paired-End DNA Reads to a Reference Genome

Anish Man Singh Shrestha and Martin C. Frith

Computational Biology Research Center, AIST, Tokyo, Japan

Many high-throughput sequencing experiments produce paired DNA reads. Paired-end DNA reads provide extra positional information that is useful in reliable mapping of short reads to a reference genome as well as in downstream analyses of structural variations. We present a new probabilistic framework to predict the alignment of paired-end reads to a reference genome. Using both simulated and real data, we compare the performance of our method against six other read-mapping tools that provide a paired-end option. We show that our method provides a good combination of sensitivity, error rate, and computation time, especially in more challenging and practical cases such as when the reference genome is incomplete or unavailable for the sample, or when there are large variations between the reference genome and the source of the reads. An open-source implementation of our method is available as part of Last, a multi-purpose alignment program freely available at http://last.cbrc.jp.