DUK – A Fast and Efficient Kmer Matching Tool

Mingkun Li[1], Alex Copeland[1], James Han[2]

[1] Lawrence Berkeley National Laboratory
[2]Lawrence Livermore National Laboratory

February 2011

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DUK – A Fast and Efficient Kmer Matching Tool

***Mingkun Li[1]**, Alex Copeland[1] and James Han[2]

[1]Lawrence Berkeley National Laboratory

[2]Lawrence Livermore National Laboratory

US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA

Email: mli@gmail.com

Abstract

A new tool, DUK, is developed to perform matching task. Matching is to find whether a query sequence partially or totally matches given reference sequences or not. Matching is similar to alignment. Indeed many traditional analysis tasks like contaminant removal use alignment tools. But for matching, there is no need to know which bases of a query sequence matches which position of a reference sequence, it only need know whether there exists a match or not. This subtle difference can make matching task much faster than alignment. DUK is accurate, versatile, fast, and has efficient memory usage. It uses Kmer hashing method to index reference sequences and Poisson model to calculate p-value. DUK is carefully implemented in C++ in object oriented design. The resulted classes can also be used to develop other tools quickly. DUK have been widely used in JGI for a wide range of applications such as contaminant removal, organelle genome separation, and assembly refinement. Many real applications and simulated dataset demonstrate its power.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

LLNL-ABS-475246