K-mer-based host screening

Bonnie Hurwitz, Ken Youens-Clark

Abstract

K-mer-based approaches to determining sequence similiarity can be faster than traditional use of BLAST. This application uses Jellyfish to index host FASTA files, then uses the mode value (default 2) of the number of matching k-mers (default k=20) from the input sequence to determine whether the sequence is similar enough to the host sequence to be rejected. The output of the app is:

- "screened" directory containing the sequences from each file that were found to be dissimlar to the host
- "rejected" directory containing the sequence from each file that were too similar to the host
- "jf" directory containing Jellyfish indexes of each "host" file (useful for later runs with other files to skip recreating)
- "kmer" directory containing k-mers of query sequences and ".loc" file showing the number of ".kmer" lines associated to each sequence

Code is freely available at Github.

Citation: Bonnie Hurwitz, Ken Youens-Clark K-mer-based host screening. **protocols.io** dx.doi.org/10.17504/protocols.io.ehjbb4n
Published: 03 Feb 2016

Protocol

Step 1.
Upload your data to the CyVerse Data Store.

Step 2.
Login to iPlant/CyVerse Discovery Environment. Choose the "Apps" button on the left, then navigate to "Public Apps -> Experimental -> iMicrobe -> Host Screen."

Step 3.
Indicate the directory containing the host (reference) sequences and the query files you wish to screen. You can change the k-mer size (default 20) and mode minimum (default 2). Click the "Launch Analysis" and wait for an email indicating that the app has finished. See the "run-screen-host" directory for the files described above.