PCPipe: Protein clustering with SIMAP annotations

Bonnie Hurwitz, Ken Youens-Clark

Abstract

PCPipe is a protein-clustering tool. The input is a set of ORFs and a FASTA file with already clustered ORFs.

The process entails:

- Use cd-hit-2d to compare the input peptides to previously clustered proteins
- The result is a file with input proteins that clustered to existing clusters and those that did not
- Use the unclustered peptides and self-cluster them via cd-hit
- Take a representative sequence from each novel cluster, and use "blastp" to compare to SIMAP.
- Use the resulting SIMAP "feature_id" to look up the SIMAP features, merging the query results with the protein ID into a tab-delimited annotations file
- Provide the user with two cluster files and the annotations for the new clusters based on the representative sequence

Code is freely available at Github.

Citation: Bonnie Hurwitz, Ken Youens-Clark PCPipe: Protein clustering with SIMAP annotations. **protocols.io** dx.doi.org/10.17504/protocols.io.ehfbb3n
Published: 02 Feb 2016

Protocol

Step 1.
Login to the iPlant/CyVerse "Discovery Environment." Choose the "Apps" button on the left, then navigate to "Public Apps -> Experimental -> iMicrobe -> PCPipe." Click on the "PCPipe" app or highlight it and choose "Apps -> Use App..." from the "Apps" menu bar.

Step 2.
Upload your data into the Data Store.

Step 3.
Indicate the directory containing the proteins/ORFs you wish to cluster along with the existing cluster file. Both files should be in FASTA format. You can alter the minimum number of members in a cluster from the default of '2' to make the clustering more stringent. Press "Launch Analysis" and wait for an email saying that the job has completed.

Step 4.
See the "pcpipe-out" directory for the following:

- "cdhit-2d-outdir" directory containing the clustered proteins and the "novel.fa" unclustered

proteins

- "cdhit-outdir" directory containing self-clustered proteins
- A file called "novel.fa" of the representative sequences from the self-clustered proteins
- A file called "blast.out" containing the "blastp" results of the "novel.fa" proteins to SIMAP
- A file called "simap-annotations.tab" showing the protein ID merged with SIMAP feature data