Title
Metagenomic gene annotation by a homology-independent approach

Permalink
https://escholarship.org/uc/item/6rj6r4mv

Author
Froula, Jeff

Publication Date
2012-08-06
I. Abstract

Fully understanding the genetic potential of a microbial community requires functional annotation of all the genes it encodes. The recently developed deep metagenome sequencing approach has enabled rapid identification of millions of genes from a complex microbial community without cultivation. Current homology-based gene annotation fails to detect distantly-related or structural homologs. Furthermore, homology searches with millions of genes are very computational intensive.

To overcome these limitations, we developed rhModeller, a homology-independent software pipeline to efficiently annotate genes from metagenomic sequencing projects. Using cellulases and carbonic anhydrases as two independent test cases, we demonstrated that rhModeller is much faster than HMMER but with comparable accuracy, at 94.5% and 99.9% accuracy, respectively. More importantly, rhModeller has the ability to detect novel proteins that do not share significant homology to any known protein families.

As ~50% of the 2 million genes derived from the cow rumen metagenome failed to be annotated based on sequence homology, we tested whether rhModeller could be used to annotate these genes. Preliminary results suggest that rhModeller is robust in the presence of missense and frameshift mutations, two common errors in metagenomic genes. Applying the pipeline to the cow rumen genes identified 4,990 novel cellulases and 8,196 novel carbonic anhydrase candidates.

In summary, we expect rhModeller to dramatically increase the speed and quality of metagenomic gene annotation.

II. Motivation

- Can not annotate genes lacking functionally characterized homologues
- Can not annotate genes with only structural similarity
- Can not identify novel or new enzymes
- The annotation may not be correct if the best match is not a true ortholog

Goal: Develop a high through-put gene annotation method for metagenomic data

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof; or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof; or The Regents of the University of California.