Title

FASTERp: a Feature Array Search Tool for Estimating Resemblance of Protein Sequences

Authors

Macklin, Derek
Egan, Rob
Wang, Zhong

Publication Date

2014-03-18

FASTERp: A Feature Array Search Tool for Estimating Resemblance of Protein Sequences

Derek N. Macklin*, Rob Egan, Zhong Wang

Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598, USA;

***** dmacklin@lbl.gov

March 2014

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

DISCLAIMER

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.

FASTERp: A Feature Array Search Tool for Estimating Resemblance of Protein Sequences

Derek Macklin*, Rob Egan, Zhong Wang
Joint Genome Institute, Walnut Creek, CA, USA
*dmacklin@lbl.gov

Metagenome sequencing efforts have provided a large pool of billions of genes for identifying enzymes with desirable biochemical traits. However, homology search with billions of genes in a rapidly growing database has become increasingly computationally impractical. Here we present our pilot efforts to develop a novel alignment-free algorithm for homology search. Specifically, we represent individual proteins as feature vectors that denote the presence or absence of short kmers in the protein sequence. Similarity between feature vectors is then computed using the Tanimoto score, a distance metric that can be rapidly computed on bit string representations of feature vectors. Preliminary results indicate good correlation with optimal alignment algorithms (Spearman *r* of 0.87, ~1,000,000 proteins from Pfam), as well as with heuristic algorithms such as BLAST (Spearman *r* of 0.86, ~1,000,000 proteins). Furthermore, a prototype of FASTERp implemented in Python runs approximately four times faster than BLAST on a small scale dataset (~1000 proteins). We are optimizing and scaling to improve FASTERp to enable rapid homology searches against billion-protein databases, thereby enabling more comprehensive gene annotation efforts.