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

FASTERp: A Feature Array Search Tool for Estimating Resemblance of protein sequences

Derek N. Macklin, Rob Egan, Zhong Wang

The Joint Genome Institute, Walnut Creek, CA, 94598

How can we efficiently perform homology search against billions of protein sequences?

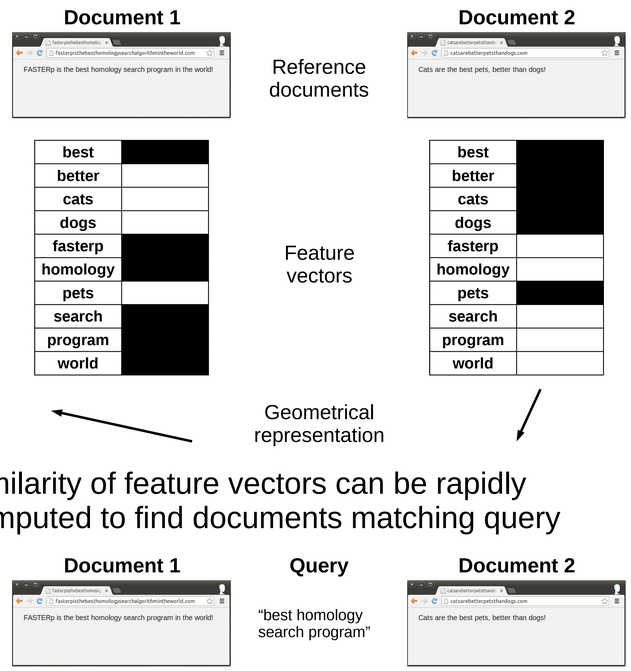
HYPOTHESIS

We can represent protein sequences as feature vectors and rapidly compute vector similarity.

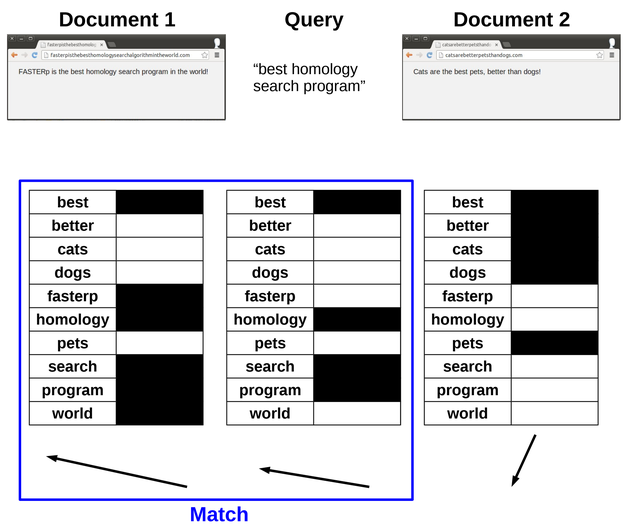
RELATED WORK

Search engines query against trillions of URLs

They represent documents as a feature vector using a “bag of words” model

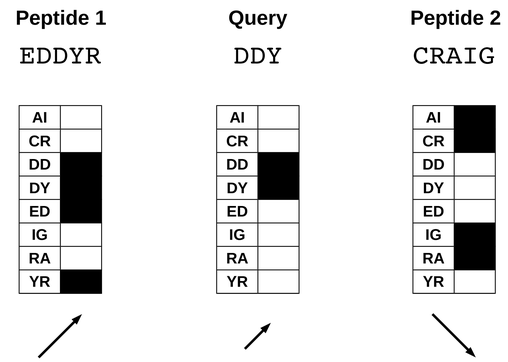


Similarity of feature vectors can be rapidly computed to find documents matching query

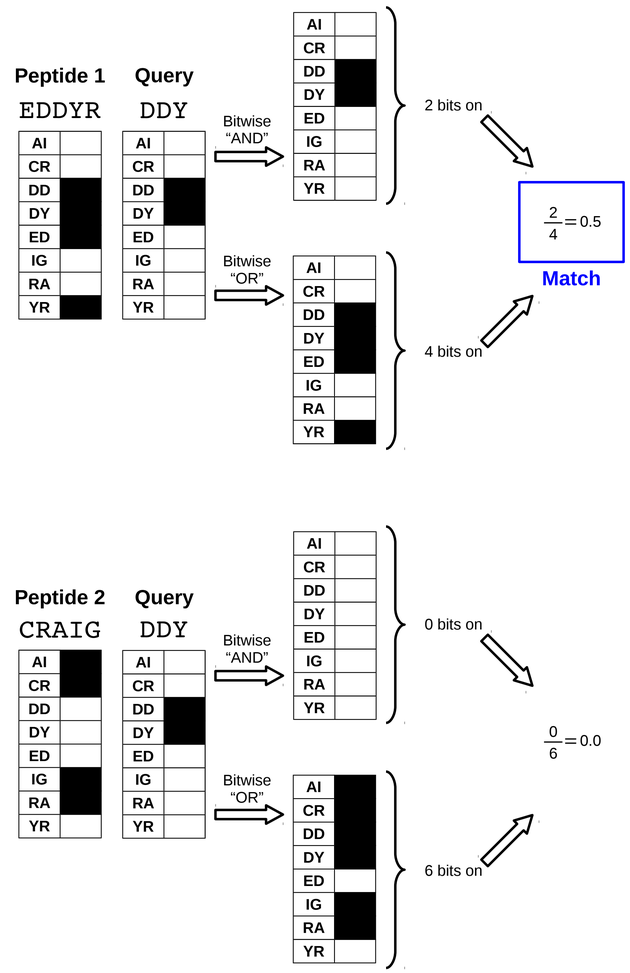


METHODS

Use short kmers as features

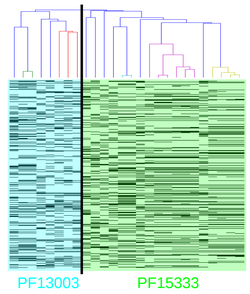


Compute similarity using the Tanimoto score

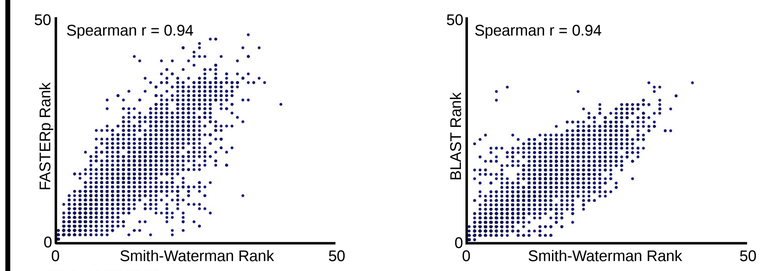


RESULTS

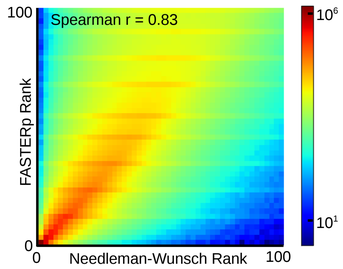
Feature vectors naturally cluster protein families



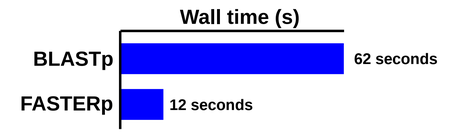
FASTERp comparable to BLAST on small data set



FASTERp performs well on larger data set



Unoptimized prototype already faster than BLAST



FUTURE WORK

Tune algorithm parameters to improve accuracy

Index database of feature vectors to reduce search time

Efficiently implement Tanimoto computations to reduce search time

DNM is supported by a Computational Science Graduate Fellowship from the Department of Energy which is provided under grant number DE-FG02-97ER25308.
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.