Using hidden Markov model emission probabilities as a general feature for the taxonomic classification of sequences

January 14, 2016

Abstract

Abstract.

1 Introduction

1.1 Feature extraction and the bias variance tradeoff

1.2 The challenge of highly divergent sequences

1.3 Homology and compositional based methods

1.4 Learning methods

2 Methods

2.1 Genelearn - modular software

2.2 Reftree- a search method for taxonomically structured data

2.3 Kmer feature extraction

2.4 Emission probability feature extraction

2.5 combining homology into composition

2.6 Learning algorithms ? Logistic regression/ SVM

2.7 GraphLab and scikit learn

2.8 Precision recall calculations

3 Results

3.1 Precision recall of Kmer vs Genemark vs Genemark + kmer for viruses and multiclass

3.2 Test of kmer length

3.3 F1 vs length of contig

3.4 Taxon level (supplemental data)

3.5 Solver comparison (supplemental)

3.6 Real metagenomic data: RdRP containing contigs

4 Discussion

4.1 The importance of feature selection (sparse vs dense, information content multiple sources of information)

4.2 Feature selection and signal to noise ratio

4.3 Homology free methods for highly divergent samples