Supplementary Data - Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification

Ivan Borozan, Stuart Watt and Vincent Ferretti

December 2014

0.1 Tables

Fragment length	PhymmBL accuracy (%)	CSSS accuracy (%)	CSSS k-mer size	CSSS (BLAST scores alone) accuracy (%)	PhymmBL (BLAST scores alone) accuracy (%)
Full genomes	86.56 ± 2.19	91.43 ± 0.99	3	82.95 ± 1.35	57.66 ± 0.93
1000	68.90 ± 1.78	70.02 ± 2.01	4	66.93 ± 1.55	31.37 ± 1.30
500	57.28 ± 2.09	63.02 ± 1.49	4	59.98 ± 1.52	26.83 ± 1.75
100	29.79 ± 1.66	35.94 ± 3.31	3	42.06 ± 1.59	19.0 ± 1.88

Table 4: Shows the classification accuracy (see eq.16) for Dataset I obtained with the CSSS (1-NN classifier) and PhymmBL models when predicting 147 different viral genera across 266 viral DNA sequences as a function of the viral fragment length. The optimum value of the k-mer parameter used by the CSSS model is indicated in column 4. Columns 5 and 6 indicate the classification accuracy obtained by each model when using BLAST scores alone.

Phylum	PhymmBL accuracy (%)	CSSS accuracy (%)	CSSS k-mer size	CSSS (BLAST scores alone) accuracy (%)	PhymmBL (BLAST scores alone) accuracy (%)
Euryarchaeota	81.14	87.03	4	32.30	0.60
Nitrospirae	97.67	96.66	4	88.77	69.07

Table 5: Shows the classification accuracy (see eq.16) for Dataset II obtained with the CSSS (1-NN classifier) and PhymmBL models when predicting the phyla for 20907 reads (with an average of 759bp in read length) belonging to Leptospirillum sp. groups II and III genomes (18579 reads) and Ferroplasma acidarmanus genome (2328 reads). The optimum value for the k-mer parameter used by the CSSS model is indicated in column 4. Columns 5 and 6 indicate the classification performance of the BLAST scores alone as implemented in CSSS and PhymmBL models.

Fragment size	CSSS(BLAST scores alone)(%)	CSSS(JSD scores alone) (%)	CSSS(ED scores alone) (%)	CSSS(CB scores alone) (%)	PhymmBL (Phymm(IMMs) scores alone) (%)	PhymmBL(BLAST scores alone) (%)
Full genomes	82.95 ± 1.35	80.01 ± 1.65	79.98 ± 1.39	69.99 ± 1.50	85.74 ± 2.55	57.66 ± 0.93
1000	66.93 ± 1.55	55.25 ± 2.15	54.21 ± 1.30	< 10	67.92 ± 2.04	31.37 ± 1.30
500	59.98 ± 1.52	42.79 ± 1.58	41.25 ± 1.66	< 10	55.78 ± 2.38	26.83 ± 1.75
100	42.06 ± 1.59	14.40 ± 2.09	13.57 ± 1.24	< 10	24.36 ± 1.86	19.0 ± 1.88

Table 6: Shows the classification accuracy (see eq.16) obtained with individual similarity/distance measures used by the CSSS (1-NN classifier) and PhymmBL models for predicting 147 different viral genera across 266 viral DNA sequences as a function of the viral fragment length. The values of the k-mer parameter used by the JSD (see eq.5) and ED (see eq.3) similarity measures are identical to those presented in Table.4.

0.2 Figures

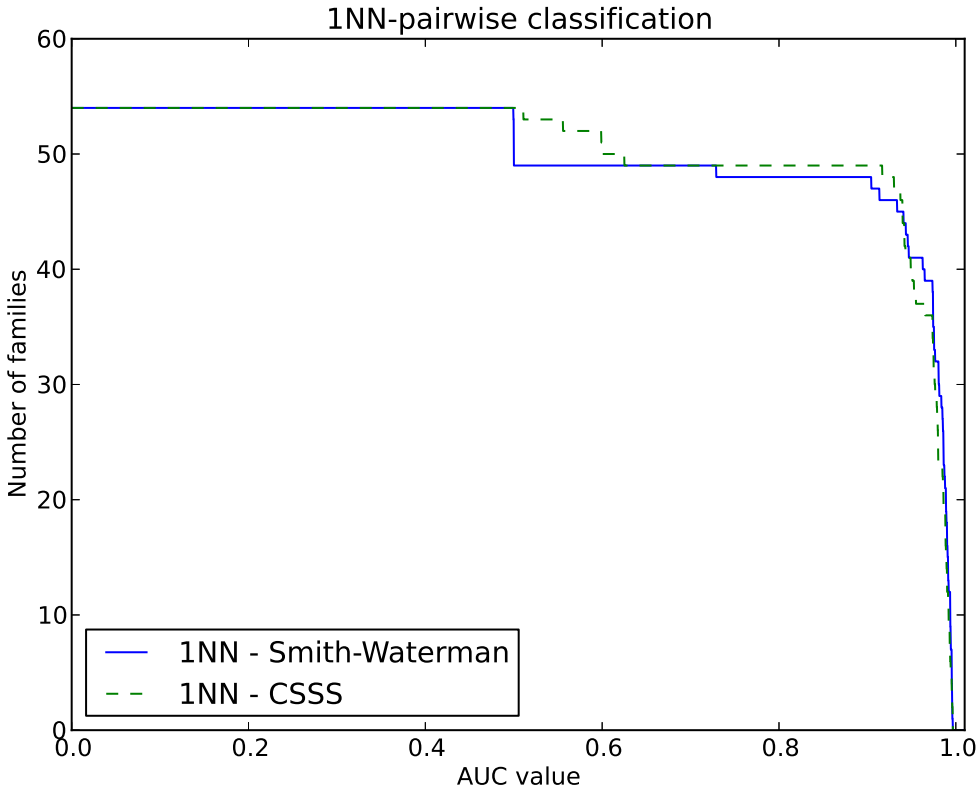
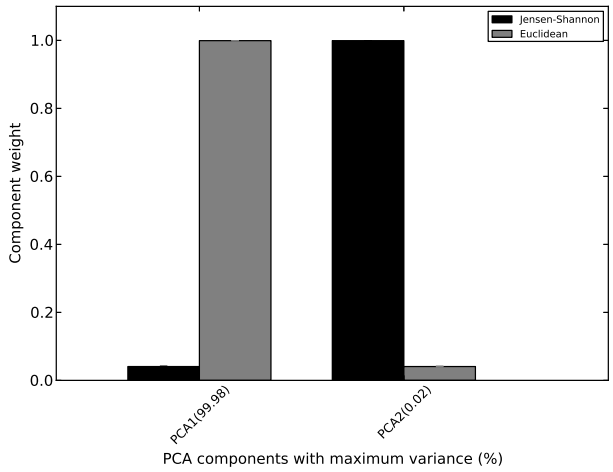
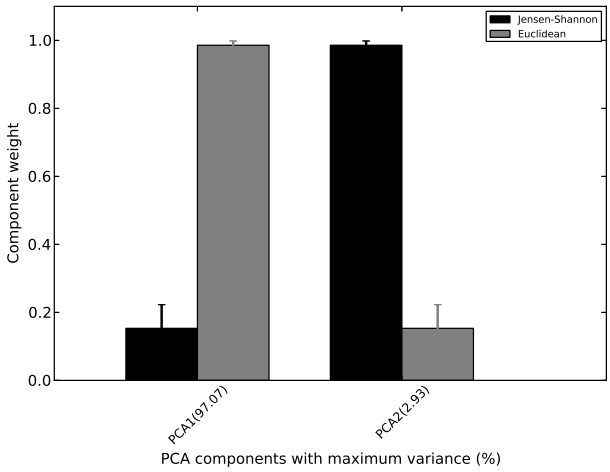


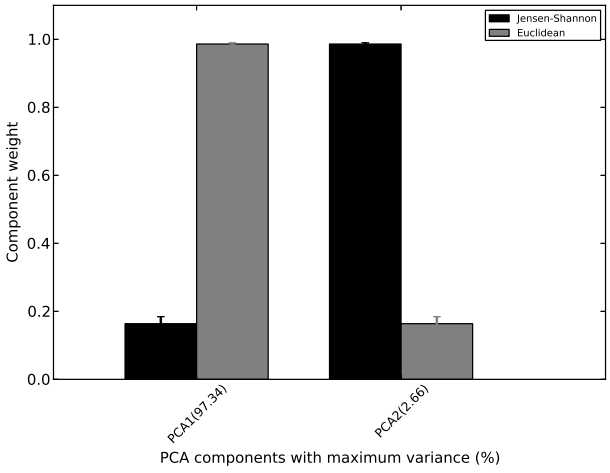
Figure 2: Shows relative classification performance of the CSSS model with the 1-NN classifier and the best performing classifier (i.e. Smith-Waterman p-values with the 1-NN classifier) presented in Kocsor et *al., 2006* on Dataset III used in this study. The graph plots the total number of families for which the integral of the ROC curve (AUC) exceeds a score threshold indicated on the x-axis. A higher curve indicates a more accurate classification performance.



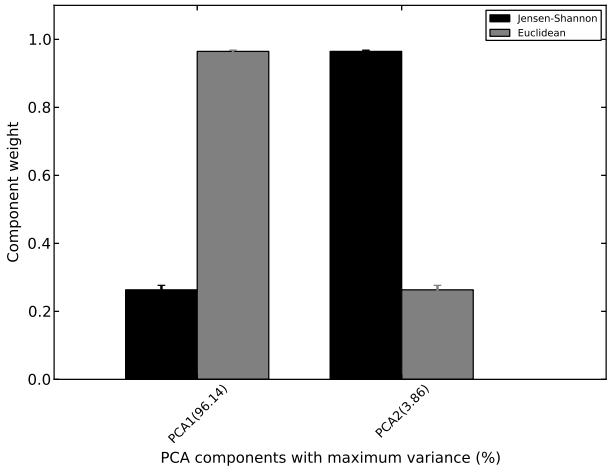
(a) Viral fragment size : full genomes



(b) Viral fragment size : 1000bp



(c) Viral fragment size : 500bp



(d) Viral fragment size : 100bp

Figure 3: Shows results of the PCA analysis of similarity scores obtained using viral genomes from Dataset I. On the x-axis are shown the first two components and on the y-axis relative weights of each component. The bar plots show that the first component (PCA1) is mostly associated with the ED measure while the second component (PCA2) is mostly associated with the JSD measure independently of the viral fragment length.

0.3 Set of parameters used to run different models/algorithms

NBC (default):
To build the database:
>countncbi genomes_training_directory 15
To score:
>score -a reads_test.fasta -r 15 -j genomes_directory
Kraken (default):
To build the database:
>kraken-build –add-to-library training.fasta –db genomeDB
>kraken-build –build –db genomeDB
To run Kraken:
>kraken –preload –db genomeDB reads_test.fasta –output results_kraken.txt
RAIphy (default):
To build the database:
>raiphy -m 2 -i training.fasta
To run RAIphy:
>raiphy -i reads_test.fasta -d defaultDb -m 1 -o results_RAIphy.txt
PAUDA (default):
To build the database:
>pauda-build training.fasta paudaDB
To run PAUDA:
>pauda-run –slow reads_test.fasta results_pauda.blastx paudaDB
PhymmBL (default):
To build the database:
>customGenomicData.pl Config.txt
To run PhymmBL:
>scoreReads.pl reads_test.fasta
PhyloPythiaS:
Please see the instruction on the following webpage:
http://phylopythias.cs.uni-duesseldorf.de/index.php?phase=wait