Supplemental Information for Use of DFT Distance Metrics for classification of SARS-CoV-2 Genomes

Micah Thornton and Monnie McGee

December 8, 2021

This document contains original research associated with the primary manuscript "Use of DFT Distance Metrics for Classification of SARS-CoV-2 Genomes". The enclosed results are referenced in the body of the manuscript, and offer further information about the primary results disclosed in the manuscript. This supplement is a PDF file that contains two sections, one with supplemental tables and associated descriptions, and one with supplemental figures and descriptions.

1 Supplemental Tables

Tables which contain further information, or are captured here for historical documentation of the evolution of the manuscript, are present in this section.

1.1 Description of Distance Methods Compared

Below is a table which enumerates the classical post alignment techniques compared in the experimental section of the manuscript. The abbreviations used in Figure 3 of the manuscript are listed in column 2 of Table S1 as a guide to the methods being compared. In the analysis section of the manuscript there is a comparison of the pairwise distances that were produced from a set of distance calculation methods. The methods applied to the data set were from the R ape package [Paradis and Schliep, 2019]. These are described and the authors referenced and attributed in Table S1.

Table S1: Kinds of post-MSA Phylogenetic Comparisons

Method Name	**Abbr.**	**Authors**	**Description**
Transversions Count	TV	-	A count of the number of transversions, that is A or G to C or T mismatches in aligned sequences
Galatier-Gouy	GG96	Nicolas Galtier Manolo Gouy	A substitution model based approach which allows a time-varying coefficient for specific nucleotides. [Galtier and Gouy, 1995]
Transitions Count	TS	-	A count of the number of transitions counted, that is A to G or C to T mismatches in aligned sequences
Barry-Hartigan	BH87	Daniel Barry John A Hartigan	A distance that is asymetric, and treats each possible transition/transversion differently [Barry and Hartigan, 1987]
Paralinear distance	paralin	James A. Lake	Uses an assymetric substitution model such as that used in the BH87 distance. [Lake, 1994]

MEGA	TN93	Koichiro Tamura and Masatoshi Nei	Specific kinds of transitions and transversions are weighted differently in this approach [Kumar *et al.*, 1994] [Tamura and Nei, 1993]
Jukes-Cantor	JC69	Thomas C. Jukes Charles R. Cantor	An early technique which uses the mutations to determine the distance by a substitution model, this is a function of the P-distance (raw/N) [Jukes *et al.*, 1969]
Generalized Jukes-Cantor	F81	Joseph Felsenstein	A theoretical technique that is the generalization of the Jukes-Cantor procedure in that it allows for different rates for each possible mismatch. [Felsenstein, 1981]
P-distance	raw	-	Proportion of sites that differ (sum of transversions and transitions) same as N.
P-distance	N	-	Proportion of sites that differ (sum of transversions and transitions) same as raw.
Kimura 3 parameteri distance	K81	Motoo Kimura	Jukes Cantor distance updated to allow different rates of 2 kinds transversions and transitions. [Kimura, 1981]
Kimura 2 parameter distance	K80	Motoo Kimura	Jukes Cantor distance updated to allow different rates of transversions and transitions. [Kimura, 1980]
Phylip	F84	Joseph Felsenstein and Gary A Churchill	A Markov-model based approach to phylogenetic distance calculations based on likely substitution states. [Felsenstein and Churchill, 1996]
Tamura (1992)	T92	Koichiro Tamura	assumes bases occur with unequal proportions, and allows variable rates in modeling procedure [Tamura, 1992].
K-mer distance	fivermers	Multiple	Euclidean distance computed between frequency vectors for the k-mers where $k \in \{1, 2, 3, 4, 5\}$.
Log Determinant	logdet	Peter J Lockhart and Michael A Steel	Another generalization of the BH procedure. [Lockhart *et al.*, 1994]
DFT Power Spectral Distances	DFTPS	Changchuan Yin & Stephen S-T. Yau	Fourier transforms taken in the described manner produce power spectra which are scaled prior to the computation of Euclidean distances [Yin, 2020].
Insertion Deletion Count	indel	-	Counts mismatches among aligned sequences where one has a gap character (-).
Chained Insertion Deletion Count	indelblock	-	Counts mismatches among aligned sequences where one has one or more gap character(s) (*-*).

Note, most of the data in Table S1 comes from the thorough documentation of the `ape` *package in R [Paradis and Schliep, 2019] [R Core Team, 2021].*

1.2 Accuracy Metric Tables for Classification Procedures

In the manuscript, we utilize the genomic Fourier power spectra to classify viral sequences by their originating location. The accuracy table included here as Table S2 appeared in the original manuscript as Table 3. It was pointed out by reviewers that the overall accuracy metric which is reported in this situation is not the best metric to utilize due to the imbalance in the dataset (for instance, always guessing 'Europe' would produce an accuracy of 48%.). The manuscript has been updated to reflect this, and instead of accuracy, the F-measure is reported for each classifier.

Table S2: Five-fold Mean Cross Validation Accuracy and Binomial Confidence Intervals for classifiers trained with k-mer frequency vectors, and DFT power spectra vectors separately, for these approaches, all of the Fourier coefficients were used

Supervised Learner	**k-mer Vectors (Interval)**	**DFT Power Spectra (Interval)**
Naive Bayes	0.42448 (0.41126,0.43770)	**0.59341 (0.58027,0.60656)**
K-Nearest Neighbors (10)	0.72226 (0.71028,0.73425)	**0.77595 (0.76479,0.78710)**
Random Forest (500)	0.65140 (0.63865,0.66415)	**0.80530 (0.79470,0.81589)**
Neural Network(1 HL - 30 N)	0.50537 (0.49199,0.51875)	**0.57981 (0.56661,0.59302)**
Support Vector Machine	0.68790 (0.67551,0.70030)	0.71152 (0.69940,0.72365)

Note that Table S2 (Previously Table 3 in the Manuscript) reports the overall five-fold cross validation accuracy of each of the classification procedures. Due to the imbalance of genomes assigned to geographic regions in the dataset, does not accurately display the superiority of the power spectra in classification tasks. Tables S3 and S4 contain the complete confusion matrices as well as class specific precision, recall, and F-measure for both the k-mer frequency vector method and the power spectra method.

2 Supplemental Figures

This section contains supplemental figures for the analysis that was performed and documented in the original manuscript.

2.1 Unsupervised Learning/ Clustering Metrics

In the original manuscript we supply visualizations using T-SNE of the data in its power spectral representation and k-mer frequency representation, in addition to these, T-SNE plots are produced for two classical distances. Besides T-SNE plots which are displayed in Figure 1 of the primary manuscript, we also produced UMAP/PCA plots for the PS/k-mer methods which we show here in Figure S1.

Additionally, it was suggested that the validity of clusters formed from the power spectra should be assessed. We perform this analysis by inspecting the K-means clustering procedure results (using 25 different starting positions). Figure S2 A shows that the sum of squared error exhibits a knee with 8 clusters, which validates the use of 8 geographical regions for supervised learning. Figure S2 B shows a cluster plot for K-means with $k = 8$. The cluster plot does not show very clear delineations of clusters due to the imbalance in the data. From the silhouette plot (Figure S2 C), we can see that the clusters formed appear to be internally consistent with a few notable exceptions. For example, clusters 3 and 4 tend to blend together more than others. There was also one very small cluster found with only 4 elements, which indicated high internal

Table S3: Classification Results for Classification with various classifiers trained with the first five k-mer frequency vectors

	True Submitting Region (Coarsened from 70 smaller submitting Locations)								Accuracy Metrics for Classifiers		
Classifier Predicted Region	**West Asia**	**East Asia**	**Europe**	**Africa**	**North America**	**South America**	**Oceania**	**Middle East**	**Recall**	**Precision**	**F1-Score**
Naive Bayes											
West Asia	54	18	47	0	3	12	0	28	0.5143	0.3333	0.4045
East Asia	13	95	114	9	7	5	7	9	0.3696	0.3668	0.3682
Europe	15	27	262	4	3	6	2	22	0.3864	0.7683	0.5142
Africa	1	9	59	16	1	2	2	3	0.4571	0.1720	0.2500
North America	5	23	46	1	24	7	2	1	0.5714	0.2202	0.3179
South America	2	55	47	1	2	47	9	1	0.5281	0.2866	0.3715
Oceania	6	19	49	3	0	3	16	10	0.4211	0.1509	0.2222
Middle East	9	11	54	1	2	7	0	79	0.5163	0.4847	0.5000
KNN Classifier											
West Asia	67	8	18	0	2	5	3	10	0.6381	0.5929	0.6147
East Asia	8	168	43	4	6	4	8	5	0.6537	0.6829	0.6680
Europe	13	53	551	11	13	16	7	16	0.8127	0.8103	0.8115
Africa	0	4	10	14	0	0	1	1	0.4000	0.4667	0.4308
North America	5	4	8	1	21	2	0	2	0.5000	0.4884	0.4941
South America	5	7	14	0	0	55	3	0	0.6180	0.6548	0.6358
Oceania	1	9	11	3	0	2	16	2	0.4211	0.3636	0.3902
Middle East	6	4	23	2	0	5	0	117	0.7647	0.7452	0.7548
Random Forest											
West Asia	50	8	8	0	5	4	1	6	0.4762	0.6098	0.5348
East Asia	11	191	125	15	7	14	16	8	0.7432	0.4935	0.5932
Europe	25	48	484	15	21	17	8	19	0.7139	0.7598	0.7361
Africa	0	0	1	2	0	0	0	0	0.0571	0.6667	0.1053
North America	1	1	12	0	9	0	1	0	0.2143	0.3750	0.2727
South America	3	5	6	0	0	51	5	2	0.5730	0.7083	0.6335
Oceania	0	0	1	2	0	1	5	0	0.1316	0.5556	0.2128
Middle East	15	4	41	1	0	2	2	118	0.7712	0.6448	0.7024
Neural Network											
West Asia	0	1	9	2	0	0	1	1	0.0000	0.0000	0.0000
East Asia	9	148	69	10	6	10	13	7	0.5759	0.5441	0.5595
Europe	32	72	525	17	18	24	10	35	0.7743	0.7162	0.7442
Africa	12	11	15	2	0	2	1	93	0.0571	0.0147	0.0234
North America	6	4	9	1	14	3	0	1	0.3333	0.3684	0.3500
South America	1	4	7	2	1	0	8	1	0.0000	0.0000	0.0000
Oceania	4	12	17	0	0	46	4	2	0.1053	0.0471	0.0650
Middle East	41	5	27	1	3	4	1	13	0.0850	0.1368	0.1048
Support Vector Machine											
West Asia	49	3	5	0	1	5	1	11	0.4667	0.6533	0.5444
East Asia	5	140	38	6	5	7	10	6	0.5447	0.6452	0.5907
Europe	44	104	611	24	23	29	18	43	0.9012	0.6819	0.7764
Africa	0	1	0	3	0	0	0	0	0.0857	0.7500	0.1538
North America	0	0	5	0	13	0	0	0	0.3095	0.7222	0.4333
South America	0	5	6	0	0	48	4	0	0.5393	0.7619	0.6316
Oceania	0	2	2	1	0	0	5	1	0.1316	0.4545	0.2041
Middle East	7	2	11	1	0	0	0	92	0.6013	0.8142	0.6917

consistency among these four samples and high distinction from others. Figure S2 displays the results of the k-means clustering analysis.

References

[Barry and Hartigan, 1987] Barry, D. and Hartigan, J. A. 1987. Statistical analysis of hominoid molecular evolution. *Statistical Science* 2, 191–207.

[Felsenstein, 1981] Felsenstein, J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution* 17, 368–376.

[Felsenstein and Churchill, 1996] Felsenstein, J. and Churchill, G. A. 1996. A hidden markov model approach to variation among sites in rate of evolution. *Molecular biology and evolution* 13, 93–104.

Table S4: Classification Results for Classification with various classifiers trained with the complete scaled power spectra

	True Submitting Region (Coarsened from 70 smaller submitting Locations)								Accuracy Metrics for Classifiers		
Classifier Predicted Region	**West Asia**	**East Asia**	**Europe**	**Africa**	**North America**	**South America**	**Oceania**	**Middle East**	**Recall**	**Precision**	**F1-Score**
Naive Bayes											
West Asia	67	32	35	17	18	2	1	4	0.6381	0.3807	0.4769
East Asia	17	124	89	7	7	16	5	7	0.4825	0.4559	0.4688
Europe	1	50	460	3	7	9	9	31	0.6785	0.8070	0.7372
Africa	3	7	1	5	0	0	1	1	0.1429	0.2778	0.1887
North America	0	16	28	0	6	2	2	6	0.1429	0.1000	0.1176
South America	0	0	1	1	0	57	0	1	0.6404	0.9500	0.7651
Oceania	4	19	20	0	1	2	15	8	0.3947	0.2174	0.2804
Middle East	13	9	44	2	3	1	5	95	0.6209	0.5523	0.5846
KNN Classifier											
West Asia	74	5	15	5	1	1	3	4	0.7048	0.6852	0.6948
East Asia	9	203	35	4	2	7	9	8	0.7899	0.7329	0.7603
Europe	15	29	571	10	7	13	5	15	0.8422	0.8586	0.8503
Africa	1	3	8	13	0	0	1	0	0.3714	0.5000	0.4262
North America	0	3	7	0	29	1	0	9	0.6905	0.5918	0.6374
South America	1	3	11	2	0	64	0	2	0.7191	0.7711	0.7442
Oceania	1	7	6	1	0	1	17	2	0.4474	0.4857	0.4658
Middle East	4	4	25	0	3	2	3	113	0.7386	0.7338	0.7362
Random Forest											
West Asia	73	1	11	5	1	0	0	4	0.6952	0.7684	0.7300
East Asia	9	203	21	4	2	11	13	7	0.7899	0.7519	0.7704
Europe	18	39	626	13	9	17	11	24	0.9233	0.8269	0.8725
Africa	0	0	0	10	0	0	0	0	0.2857	1.0000	0.4444
North America	0	1	0	0	27	0	0	4	0.6429	0.8438	0.7297
South America	0	1	3	1	0	61	0	0	0.6854	0.9242	0.7871
Oceania	1	5	1	0	0	0	11	0	0.2895	0.6111	0.3929
Middle East	4	7	16	2	3	0	3	114	0.7451	0.7651	0.7550
Neural Network											
West Asia	2	4	5	8	0	0	2	0	0.0190	0.0952	0.0317
East Asia	4	185	28	5	6	9	14	14	0.7198	0.6981	0.7088
Europe	24	49	591	9	10	17	8	25	0.8717	0.8063	0.8377
Africa	6	7	20	3	4	1	3	102	0.0857	0.0205	0.0331
North America	0	1	6	2	20	1	0	2	0.4762	0.6250	0.5405
South America	3	2	6	1	0	1	9	1	0.0112	0.0435	0.0179
Oceania	1	4	10	0	0	58	0	1	0.0000	0.0000	0.0000
Middle East	65	5	12	7	2	2	2	8	0.0523	0.0777	0.0625
Support Vector Machine											
West Asia	73	0	26	17	17	1	0	2	0.6952	0.5368	0.6058
East Asia	11	164	21	6	6	10	7	10	0.6381	0.6979	0.6667
Europe	8	83	585	7	10	20	17	42	0.8628	0.7578	0.8069
Africa	0	0	0	2	0	0	0	0	0.0571	1.0000	0.1081
North America	0	2	0	0	5	0	0	0	0.1190	0.7143	0.2041
South America	0	1	3	1	0	57	0	0	0.6404	0.9194	0.7550
Oceania	0	0	1	0	0	0	9	0	0.2368	0.9000	0.3750
Middle East	13	7	42	2	4	1	5	99	0.6471	0.5723	0.6074

[Galtier and Gouy, 1995] Galtier, N. and Gouy, M. 1995. Inferring phylogenies from dna sequences of unequal base compositions. *Proceedings of the National Academy of Sciences* 92, 11317–11321.

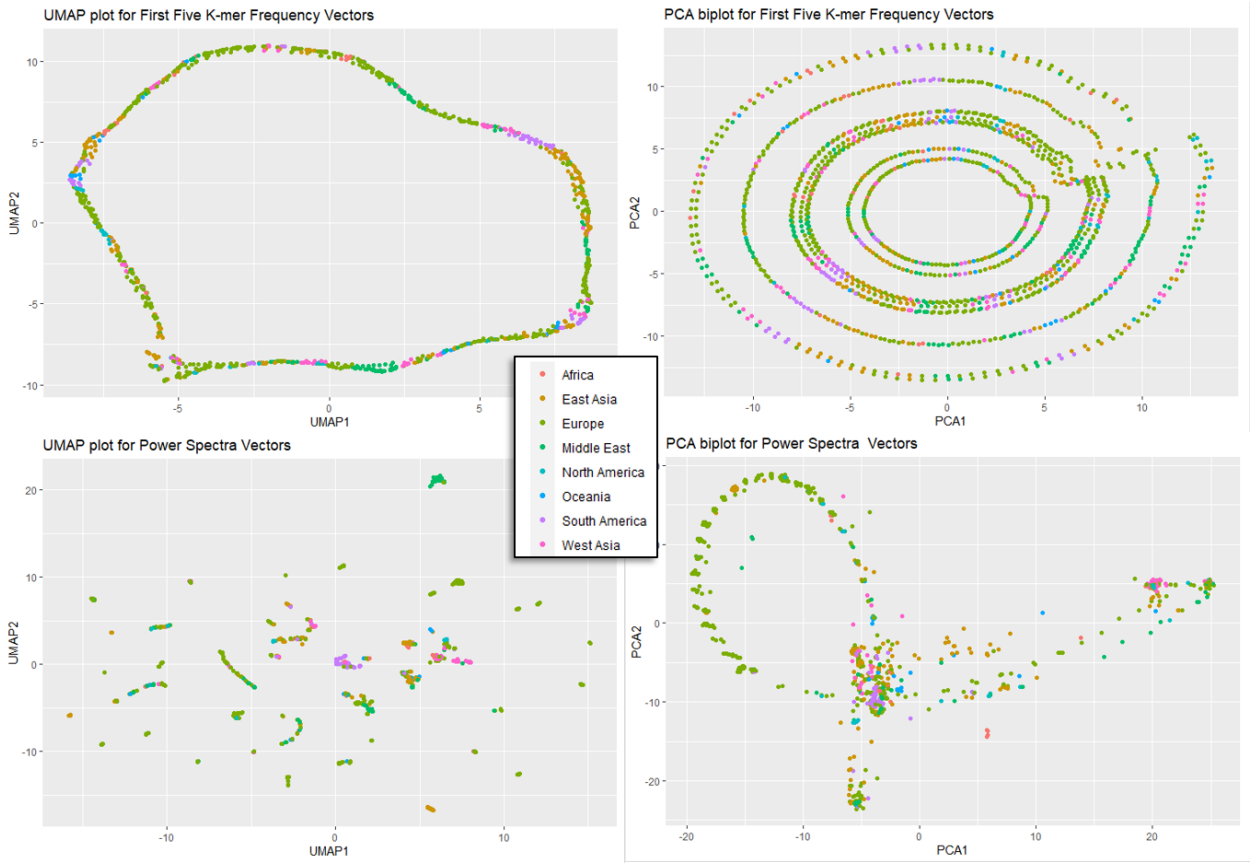
[Jukes *et al.*, 1969] Jukes, T. H., Cantor, C. R., *et al.* 1969. Evolution of protein molecules. *Mammalian protein metabolism* 3, 21–132.

[Kimura, 1980] Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* 16, 111–120.

[Kimura, 1981] Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences* 78, 454–458.

[Kumar *et al.*, 1994] Kumar, S., Tamura, K., and Nei, M. 1994. Mega: molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics* 10, 189–191.

Figure S1: UMAP and PCA visualizations for the k-mer and Power Spectral Representations of 1,397 SARS-CoV-2 viral genomes



[Lake, 1994] Lake, J. A. 1994. Reconstructing evolutionary trees from dna and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences* 91, 1455–1459.

[Lockhart *et al.*, 1994] Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular biology and evolution* 11, 605–612.

[Paradis and Schliep, 2019] Paradis, E. and Schliep, K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528.

[R Core Team, 2021] R Core Team 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Tamura, 1992] Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and g+ c-content biases. *Mol Biol Evol* 9, 678–687.

[Tamura and Nei, 1993] Tamura, K. and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular biology and evolution* 10, 512–526.

[Yin, 2020] Yin, C. 2020. Phylogenetic analysis of DNA sequences or genomes by Fourier transform.

Figure S2: Validation of Produced Clusters (by K-means K=8). (A) Internal Sum of Squared Error by Number of Clusters (knee indicates optimum at K=8) (B) Cluster Plot of clusters produced by K-Means (with K=8) - from Factoextra package in R (C) Silhouette Plot for K-Means clustering with K=8.