Title.

(Draft)
Comparisons of Distance Measures between 125 Fig Cultivar SSR Profiles

Authors.

Richard Frost, 11/4/2020

Affiliations.

Frost Concepts, Vista CA USA
http://tangentvectors.org/

Author Summary.

Richard Frost is an applied mathematician (M.S. 1990, B.S. Math, Physics 1987) with practical and pedagogical experience in numerical analysis, scientific computing, and horticulture. He currently maintains a collection of cultivars from 50 different species of fruiting perennial plants in Vista CA.

Abstract.

This analysis is the result of a mathematician's query into the ancestral relationships between his fig trees. Twenty-five genetic distance measures are applied to SSR genetic profile data acquired in a previous study of fig accessions at NCGR Davis. Of those, 11 measures were selected for further study. Among those rejected are 3 in common use that do not meet the basic mathematical requirements of distance. The accepted measures are analyzed for similarities, distance resolution, matches to breeding records, and likely number of clades. The general nature of nearest-neighbor (*k*-neighbor) algorithms and their effect on standard clustering software is discussed. Nearest-neighbor components are used to form initial clusters, and then bridged to produce fully connected least bridges graphs. Topological structures are presented throughout to visualize possible relationships between fig cultivars. A summary chart of findings is provided at the beginning of the Results section. Background information concerning SSR measurements and basic mathematical principles related to distance measure and graph theory are given in the Methods section. A source of example computer programs is listed at the end of the Results section.

Introduction

Identifying plant varieties is an age-old human endeavor. Historically, morphological traits were used to categorize specimens into families, genera, species, and cultivars (a plant selected from seedlings and re-propagated for its desired characteristics). In the present day it is now possible to discern differences in plant varieties via genetic measures. Some of these are “whole sequence” while others focus on subsequences termed genetic profiles or “fingerprints”. One of these latter methods utilizes genetic profiles based repeating values in the plant genome, termed SSR for “simple sequence repeats”@[1]. So for example, if someone wishes to determine if two individual apple trees are the same cultivar, they can submit leaf samples to a plant ID lab and obtain an answer. In fact for some economically important crops, databases of SSR fingerprints have been established – so a plant ID lab can sometimes also determine which apple tree cultivar(s) the leaf specimens are from@[2]. In addition to plant ID, those involved with plant breeding and germplasm repositories wish to determine the relationships among cultivars in a given collection, if not all cultivars worldwide@[3]. For this application a measure of “distance” between SSR profiles is needed, and it is helpful to have some reliable breeding records of a few individuals to establish ground truth and a scale of distance.

The data collected in an SSR profile are from a series of genetic loci deemed important for the genus or species, using values at one or more alleles per loci. These “values” can be a letter or letter sequence representing amino acid(s) particular to that allele, or a numerical value representing the quantity of the particular acid(s), or an index representation of the allele contents. From the mathematical viewpoint they are spatial in nature. Values in the form of non-numeric representations cannot be used directly for computing a distance; e.g. a distance based on lexicographic measure will have no match to the dynamics of the profiles. These values are instead transformed to the spectral domain to provide numeric measure. To capture genetic function, the frequencies can be calculated throughout the sample set per allele, or per locus, or over all alleles and loci of the sample population. For example, Table #[Ex1] shows values of a profile in the spatial domain and then the frequencies in the spectral domain per locus.

Table #[Ex1]. Example 2×5 SSR profile from a sample population of 30 individuals. In this example the spatial values are symbols of the dominant repeat type for each allele of each locus. The computed frequencies per locus reveal that CT occurred 20 times in Locus 1 but only 12 times in Locus 3.

Spatial:					
	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5
Allele 1	CT	AG	CT	GA	CA
Allele 2	CT	AAG	TC	TC	TG
Spectral:					
Allele 1	1/3	1/3	1/5	1/6	1/6
Allele 2	1/3	1/6	1/6	1/3	1/15

Although SSR profiles are relatively new, genetic distance measures date back about a century. Computations were done by hand in those early days, or with relatively meager computers until the 1980's. Complication was to be avoided in deference to computational ease. This might explain why the discipline of vector analysis established in the 1920’s@[Katz] was under-utilized at the time. As technology evolved, sample sizes and complexity have grown along with the computerization of the geneticists’ tasks. It seems that nowadays few practitioners implement distance formulas – instead

taking the output from one machine, putting it in another and running a ready-made package. This is a tremendous increase in productivity but from results in this study it appears that some of today's data sets exceed the capabilities of several "standard" formulas. Ideally, providers of software to practitioners should include automatic safe-guard tests but apparently this is not always the case.

Genetic distance measures can be roughly classified into 3 categories: dynamic, statistical, and geometric. Dynamic methods use knowledge of linkage locations of loci along a genetic sequence to produce simulations of genetic cross-overs in breeding, then analyze allele values to compute probabilities of relationships. Centimorgans are an example measure produced by dynamic simulation@[Cent]. In contrast, statistical and geometric measures do not require linkage data – which is a simplification in data acquisition, computation, and cost. Care however must be taken to determine which measures – if any, are relevant to the data. Statistical measures of genetic distance have their roots in comparing differences in populations, mostly originating in Fisher's 1930 treatise on genetic variance@[Fish]. Soergel is an example statistical measure of distance. Geometric measures and their extensions in topology use norm and norm-like measures to compute distances between spatial or spectral values. They can be subclassified into those designed for vectors and tensors. The Frobenius norm and Spectral Radius Angle are two examples of geometric measures.

Of interest in the present study are SSR profiles taken in 2010 of the *Ficus carica* (fig) and *F. palmata* (Indian fig) collection at NCGR Davis@[Aradhya]. Structurally the data are 2×15 arrays of spatial data representing the total number of repeats of the dominant type per allele of 15 loci with 2 alleles each. An example is given in Table #[Kadota].

Table #[Kadota]. Spatial SSR Profile from NCGR 2010 for *F. carica* cultivar "Kadota" with locus names across the top. Values are the total number of repeats of the dominant type per allele (nomenclature: bp).

C22F1	C24H1	C26N1	C31F1	C35H1	C37N1	LM12H1	LM14H1	LM30N1	LM36N1	M1F1	M2H1	M3N1	M4F1	M8N1
283	272	234	224	254	204	214	200	243	248	172	153	120	194	171
283	272	234	239	254	208	243	200	245	248	189	167	132	218	175

Note that the data are tensor in nature, yet few purely tensor genetic distance measures exist in the literature. As such it is a common – but dubious practice to "flatten" tensors into vectors for use in vector measures. But from the topological point of view this breaks the geometric dynamics of alleles which in turn are important in the genetic theory of additive and dominant interaction@[GT]. Practitioners might do well searching out measures from disciplines well-versed in tensor analysis. The Spectral Radius Angle is introduced here.

This study explores the mathematical properties of several genetic distance measures and their application to both spatial and spectral SSR profiles of the NCGR data. It is shown that several measures are inapplicable to the data. Of the applicable measures, further analysis is provided in terms of distance resolution and graph theory. Estimates of relationships between these cultivars are shown in nearest-neighbor topological graphs.

Although dendrograms are useful in ascertaining phylogenetic relations among organisms, in the author's opinion they do not appear applicable to relations between specific individuals. Hence topological graphs are used for visualization.