Strategies and procedures for diversity analysis

Methods of diversity analysis for different types of datasets

Utilization of diverse trait-specific germplasm to develop crop varieties with a broad genetic base is the main objective of various crop improvement programs worldwide. A large number of germplasm accessions of different crops have been conserved in different genebanks and are being characterized for morphological and agronomic traits as well as by using molecular markers. Efficient utilization of these germplasm accessions depends upon the diversity present in these collections, for which one of the approaches is to form clusters where accessions between clusters would be more diverse than the accessions within cluster. The clustering algorithms require a distance/similarity matrix between the accessions which can be calculated depending upon the nature or type of traits such as morphological and agronomic traits and/or molecular markers. A large number of distance measures are available for analyzing similarity/dissimilarity among accessions based on different traits representing different types of variables, and the selection of the most appropriate distance measure for each trait is the prerequisite for diversity analysis studies.

In genebanks, germplasm is characterized/evaluated using

A. Morphological traits
B. Agronomic traits
C. Molecular markers

A. Morphological traits: Data recorded on morphological traits, such as flower colour, pigmentation, seed colour etc. Represent discrete or categorical variables and can be grouped as:

a. Binary: presence or absence of a characteristic
b. Nominal: color or shape of a trait
c. Ordinal: a visual scale arranged to represent the intensity of a trait

B. Agronomic traits: Data recorded on agronomic traits such as plant height, 100-grain weight, yield per plant, etc. Represent continuous variables

C. **Molecular marker:** The data on molecular markers is recorded in the following two forms:

a. Binary data: presence or absence of molecular marker bands
b. Allelic data (i.e. on allele size)

We have used notations from Gower (1971) and Genstat (Payne, 2009) have been used to present the similarity matrices.

The most widely used similarity/dissimilarity measures for different types of traits have been given. The symbols *i* and *j* have been used to denote the units, or more technically, called operation taxonomic units (OTUs) between which the similarities or dissimilarity are computed.

Morphological traits:

The most commonly used measure for analyzing diversity among accessions based on morphological data is:

Simple matching coefficient (SM):

Simple matching coefficient is useful when all the categories of values carry equal information (symmetry). It measures the proportion of categories where both the OTUs are present, and can be written as:

$$S_{SM\,ij} = a/n$$

where $S_{SM\,ij}$ is the similarity between the two OTUs, i and j, a is the number of variables/traits where both the OTUs *i* and *j* both are present in the same categories, n is the total number of categories. The coefficient takes into account the joint absence of a variable.

A. Agronomic traits:

Commonly used measures for analyzing diversity among accessions based on the data recorded for agronomic traits are:

1. Euclidean distance: Most commonly used distance which examines the root of the sum of squared differences between coordinates of a pair of OTUs. Let x_{ik} and x_{jk} be the values of k^{th} character on the individuals i and j, respectively, (k=1, 2.., K) and K is the total number of traits.

Euclidean distance is computed from raw data, and not from standardized data. Between two given individuals i and j, it is calculated as:

$$d_{ij} = [\sum_{k=1}^{K} (x_{ik} - x_{jk})^2]^{1/2}$$

where d_{ij} is the Euclidean distance between them. The Euclidean distance is highly influenced by the measurement scale of the variables and varies from 0 to infinity. To adjust for the scale differences and since we use similarity which varies between 0(dissimilarity) and unity (perfect similarity) for clustering, we can use the range standardized Euclidean distance as

$$d_{ij} = [\sum_{k=1}^{K} (x_{ik} - x_{jk})^2 / r_k]^{1/2}$$

where the range $r_k = \max_i(x_{ik}) - \max_i(x_{ik})$.

2. Manhattan distance: Also known as City Block distance, an absolute distance value or a distance by car. It is simply the average difference across dimensions. It is calculated as:

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

where d_{ij} is the distance between two individuals i and j; x_{ik} and x_{jk} are the values of k^{th} character on the individuals i and j, respectively. Its range standardized version is

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}| / r_k$$

The effect of single large differences (outliers) is dampened since they are not squared.

3. Mahalanobis D^2 distance: Mahalanobis D^2 distance (Mahalanobis, 1936) is based on a multi-variate distance between two OTUs. It differs from the Euclidean distance in the sense that it takes into account the correlations between the variables and does not depend on the scale of measurement.

III. Molecular data

For allelic data

1. Euclidean
2. Roger's distance
3. Cavalli-Sforza and Edward's Chord distance

1. Euclidean distance: The Euclidean distance is used when allelic informative marker data (if allele frequencies can be determined from the molecular data) is available on a sample of accessions from the populations, the operational taxonomic units (OTUs). This distance satisfies the Euclidean property (Reif et al., 2005).

The Euclidean distance (d_E) between the two OTUs under consideration in the case of frequencies is defined as:

$$d_E = \left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} (p_{ij} - q_{ij})^2 \right]^{1/2}$$

where p_{ij} and q_{ij} are allele frequencies of the j^{th} allele at the i^{th} locus in the two OTUs, n_i is the number of alleles at the i^{th} locus, and m refers to the number of loci.

2. Roger's distance: The Roger's distance (Roger, 1972) is a modification of the above Euclidean distance (d_E) and was developed assuming no knowledge about evolutionary forces diverging the OTUs under consideration. The Roger's distance (d_R) is given by:

$$\sum_{m} \quad \sum_{n_i}$$

$$d_R = (1/m) \sum_{i=1} [(1/2) \sum_{j=1} (p_{ij} - q_{ij})^2]^{1/2}$$

where d_R is the average d_E across all loci and further by standardizing with the factor $\sqrt{1/2}$, it lies within the interval [0,1]. d_R is suitable for studying the relationship between the genetic dissimilarity of inbreds based on allelic informative marker data and the coefficient of coancestry (Malecot, 1948; Reif et al., 2005). It is appropriate to examine the assembly and validation of core collections, and uncovering of pedigree relationships among OTUs such as the detection of essentially derived varieties in plant breeding or the identification of duplicates and collection gaps in seed banks.

3. Cavalli-Sforza and Edwards' Chord distance: This measure of the genetic distance was developed by Cavalla-Sforza and Edwards (1967) to analyze blood group alleles frequencies in human populations. This distance measure, d_{CE}, is defined as:

$$d_{CE} = [(1/m) \sum_{i=1}^{m} (1 - \sum_{j=1}^{n_i} \sqrt{p_{ij} q_{ij}})]^{1/2}$$

The d_{CE} ranges from zero to one even in the case of multiple alleles, which is an advantage over d_R. This distance measure was developed by assuming that the mutation rate is small and variation in selection pressure is rapid and haphazard i.e. no constant direction in allele frequency changes. The d_{CE} is an appropriate coefficient to investigate phylogenetic relationships among populations when allelic informative marker data are available and one can assume the selective drift model (Reif et al., 2005).

For binary data

1. Jaccard
2. Dice
3. Simple matching

Allelic data can also be converted into the binary data following ALS Binary and BinGeno analysis tool and this binary data can be analyzed following the aforementioned distance measures.

1. Jaccard: Jaccard similarity coefficient (Jaccard, 1908) is concerned only with the features that have positive co-occurrence and is defined as:

$$S_J = v_{ij} / (v_{ij} + w_{ij} + x_{ij})$$

Where S_J is the Jaccard coefficient; v_{ij} is the number of bands in common between both OTUs; w_{ij} is the number of bands present in the OTU i but not in OTU j; x_{ij} is the number of bands present in the OTU j but not in the i. This avoids the use of joint absence of the OTUs in the calculation of similarity and ranges from 0 to 1. Thus, it is the measurement of asymmetric information on binary (and non-binary) variables. It measures the proportion of shared joint presences.

2. Dice: The dice binary similarity coefficient (Dice, 1945), S_D , is defined as :

$$S_D = 2v_{ij} / (2v_{ij} + w_{ij} + x_{ij})$$

The dice coefficient is similar to the Jaccard similarity coefficient but gives twice the weight to agreements (represented by $2v_{ij}$).

3. Simple matching: The simple matching binary similarity coefficient (Sneath and Sokal, 1973) is defined as:

$$S_{SM} = (v_{ij} + y_{ij}) / (v_{ij} + w_{ij} + x_{ij} + y_{ij})$$

The $S_{SM,,}$ simple matching coefficient, takes into account matches on presences as well as absences. In above, y_{ij} is the number of bands absent from both OTUs.

Of these three measures, S_J and S_D do not involve matches of the absences (y_{ij}). Therefore, if the probability of non-amplifications of bands is high and absence of bands in both OUTs cannot be interpreted as common characteristics, it would be appropriate to use S_J and S_D instead of S_{SM}.

IV. Combination of traits:

The above distance or similarity measures are defined for multiple traits. However, these traits should be of the same type. In reality, we have different types of traits which can be taken into account by using Gower's coefficient (1971) for the following combination of traits:

- Morphological (nominal) + Agronomic traits (continuous)
- Morphological (nominal) + molecular data (binary)
- Agronomic (continuous) + molecular data (binary)
- Morphological (nominal) + agronomic (continuous) + molecular data (binary)

Gower's coefficient: Proposed by Gower (1971), it permits the simultaneous use of variables of different scales of measurement in the estimation of similarity. It is defined as:

$$S_{ij} = \{ \sum_{k=1}^{n} S_{ijk} W_{ijk} \} / \sum_{k=1}^{n} W_{ijk}$$

where S_{ij} is the derived Gower's similarity coefficient by combining similarities from different traits and giving appropriate weights; S_{ijk} denotes the contribution provided by the k^{th} variable, and W_{ijk} are the weights of each feature variable which is usually 1 or 0 depending upon whether or not the comparison is valid for the k^{th} variable; if differential variable weights are specified it is the weight of the k^{th} variable or 0 if the comparison is not valid. The weights are determined arbitrary, based on unit or based on the data (calibration). The distance D_{ij} between the OTUs *i* and *j* was computed as : $D_{ij} = 1 - S_{ij}$.

The major advantage of this measure is that the coefficient has a number of appealing features beyond its ability to accommodate mixed data types. These include its metric qualities and its flexibility, in that the method can be easily modified to include matches for absences in the estimation of similarity by simply modifying the binary weighting system. However, the coefficient has little use which may be attributed to its failure to appear in any major clustering software packages.

Upadhyaya et al. (2010) have used this measure of diversity in developing mini core collection of finger millet at ICRISAT.

Cluster analysis

The different distance matrices generated by using different distance measures were used to cluster the genotypes based upon the characteristics they possess so that the individuals with similar descriptions are grouped into the same cluster (Hair et al., 1995) and the resulting clusters of individuals should then exhibit high internal (within cluster) homogeneity and high external (between cluster) heterogeneity. For clustering, among hierarchical and non-hierarchical clustering methods, hierarchical clustering methods are more commonly employed in analysis of genetic diversity in crop species. Among various agglomerative hierarchical methods, the UPGMA (Unweighted Pair Group Method using Arithmetic averages; Sneath and Sokal, 1973) and Ward's minimum variance method (Ward, 1963) are the most commonly adopted clustering agglorithms (Mohammadi and Prasanna, 2003). The different distance matrices for agronomic traits, molecular markers and combination of traits were used for clustering following UPGMA and Ward's minimum variance method in chickpea. Amongst the two methods, Ward's method was found to be more useful in clustering the chickpea accessions for all the traits. To determine the precise number of clusters, each dendrogram (constructed based upon different distance matrices) was cut at a point where the gap between two successive combination was comparatively larger, which resulted into different number of clusters for different distance matrices for different traits in Ward's method. Clustering of accessions based upon different distance matrices using traits or their combination resulted in different number of clusters. However, a relatively higher number of accessions tend to cluster together even when different matrices/methods were used.

Results of Analysis

Traits	**Distance measure**	**Remarks**
Morphological traits	Simple	Only studied distance measure for

	matching	nominal traits
Agronomic traits	Euclidean	Identified the same pair of accessions exhibiting minimum diversity but different pair of genotypes exhibiting maximum diversity.
	Manhattan	
	D^2 Mahalanobis	Takes into account the correlations of the datasets and is scale-invariant i.e. not dependent on the scale of measurement.
Molecular markers		
Allelic data	Simple Matching	The mean as well as range of diversity was reduced, so could not discriminate the pair of accessions exhibiting maximum diversity
	Euclidean	Identified the same pair of accessions exhibiting minimum and maximum diversity
	Roger's	
	Chord 67	Assumption: mutation rate is small and variation in selection pressure is rapid and haphazard i.e. no constant direction in allele frequency changes, which is not fulfilled in seed banks and plant breeding materials that have evolved due to directed selection pressure rather than rapid and haphazard changes.
	Chord 69	
Binary data	Dice	Identified the same pair of accessions having minimum and maximum diversity
	Jaccard	
	Simple matching	Based on assumption that all shared bands (both presence and absences) are taken into account irrespective of the reason why bands are absent

Jaccard is the most appropriate when the purpose of measure of similarity/dissimilarity is to indicate how similar/different the objects are with respect to attributes present (coded as 1) and to ignore the impact of attributes absent (0).		
Combination of traits		
Morphological (nominal) + Agronomic traits (continuous)	Gower's distance	Simultaneous use of variables of different scales of measurement (nominal, continuous, binary) in the estimation of similarity/dissimilarity. has the ability to accommodate mixed data types and due to its metric qualities and flexibility, it can be modified to include negative matches in the estimation of similarity by simply modifying the binary weighting system.
Morphological (nominal) + molecular data (binary)	Gower's distance	
Agronomic (continuous) + molecular (binary)	Gower's distance	
Morphological (nominal)+ agronomic (continuous) + molecular data (binary)	Gower's distance	
Cluster analysis		
Ward's minimum variance method		Found more useful in chickpea as it grouped the genotypes into defined clusters
UPGMA (Unweighted Pair Group Method using Arithmetic averages)		In chickpea, genotypes were not grouped into clusters.

The result of analysis in chickpea showing the least and the most diverse pairs of accessions and an overall statistics of diversity for individual trait and for combination of traits have been given as an example in the attached file (Table 4.2.2).

Conclusions

- **Each distance measure has its own properties and assumptions**

- **The genetical context and mathematical properties of similarity/dissimilarity measures should be given importance when choosing a measure.**
- **Different distance measures provided different estimates of mean, minimum and maximum diversity**
- **Ward's method: Useful in clustering the chickpea accessions for all the traits**
- **Different distance measures resulted into different number of clusters for different traits, however, a relatively higher number of accessions tend to cluster together even when different matrices/methods were used.**

References

- Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: Models and estimation procedures. Am. J. Hum. Genet. 19:233–257.

- Dice LR. 1945. Measures of the amount of ecologic association between species. Ecology 26:297–302.

- Gower JC. 1971. A general coefficient of similarity and some of its properties. Biometrics 27: 857-874.

- Hair JR., Anderson RE, Tatham RL, Black WC. 1995. Multivariate data analysis with readings. 4th Edition, Prentice Hall, Englewood Cliffs, NJ.

- Jaccard P. 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44:223–270.

- Malecot G. 1948. Les Mathématiques de l'Hérédite. Masson et Cie, Paris.

- Mohammadi SA., Prasanna BM. 2003. Analysis of genetic diversity in crop plants—Salient statistical tools and considerations. Crop Sci. 43:1235–1248.

- Payne RW. 2009.The Guide to GenStat® Release 12, Part 2: Statistics. VSN International, 5 The Waterhouse, Waterhouse Street, Hemel Hempstead, Hertfordshire HP1 1ES, UK

- Reif JC, Melchinger AE, Frisch M. 2005. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Science 45:1-7.

- Rogers JS. 1972. Measures of genetic similarity and genetic distance. p. 145–153. *In* Studies in genetics VII. Publ. 7213. Univ. of Texas, Austin.

- Sneath PHA, Sokal RR. 1973. Numerical taxonomy. Freeman, San Francisco, CA.

- Ward JH. Jr. 1963. Hierarchical grouping to optimize an objective function. J. Am. Statist. Assoc. 58:236-244.

- Upadhyaya H.D, Sarma NDRK, Ravishankar CR, Albrecht T, Narasimhudu Y, Singh SK, Varshney SK, Reddy VG, Singh S, Dwivedi SL, Wanyera N, Oduori COA, Mgonja MA, Kisandu DB, Parzies HK, Gowda CLL. 2010. Developing mini core collection in finger millet using multilocation data. Crop Science (Accepted).

Training workshop and identify champions

The enhanced use of plant genetic resources conserved in various genebanks in crop improvement programs depends upon the availability of the trait-specific diverse germplasm accessions. In genebanks, the germplasm is characterized/evaluated by recording data on morphological and agronomic traits and/or molecular markers. The identification of trait-specific diverse germplasm accessions depends upon the selection of the most appropriate distance measure for different types of data sets and their combinations. Hence, the knowledge and understanding of different diversity analysis methods for different types of data sets is important for the genebank scientists involved in diversity studies. Keeping this in view, a training workshop was organized to provide technical know-how for diversity analysis using data on morphological and agronomic traits and molecular markers, and hands-on training on use of softwares such as GenStat and DARwin for diversity analysis during 1-5 September, 2009 at ICRISAT, Patancheru, India. A total of 20 scientists (12 from 8 CGIAR genebanks, 6 Indian NARS, and 2 African NARS) participated in the training workshop. All the participants strengthen their knowledge and efficiency in analyzing the data and to draw appropriate and meaningful conclusions. They felt that the training workshop provided them an opportunity to learn various methods of diversity analysis and that they were greatly benefitted by the hands-on training on the use of softwares.

To enhance scope and facilitate greater participation of expertise, eight system-wide champions were identified. The champions are

1. Alexandra Jorge (ILRI, Ethiopia)
2. Chrispus O.A. Oduori (KARI, Kenya)
3. Claudia Bedoya (CIMMYT, Mexico)
4. David Tay (CIP, Peru)
5. Fida Alo (ICARDA, Syria)
6. Heiko K. Parzies (University of Hohenheim, Germany)
7. Nelson Wanyera (NASARRI, Uganda)
8. Ranjana Bhattacharjee (IITA, Nigeria).

These champions represent various continents and are engaged in diversity analysis studies in various crops. They are capable in providing solutions to the diversity analysis related queries and would contribute to the central Help Desk.

Helpdesk for recommendation and implementation- for scientists with no support in terms of software and people

A Help Desk http://220.227.242.211:9905/ has been established to facilitate system-wide common procedures of practice for diversity analysis across CG genebanks. This Help desk would serve the genebank community globally by providing basic information on various aspects of diversity analysis such as selection of appropriate similarity/dissimilarity matrices, cluster analysis, analyzing diversity using individual trait and/or combination of traits, etc.