BIOTECHNOLOGY SOCIETY OF NEPAL (BSN) E-BULLETIN
VOLUME 2, NOVEMBER 2010

Research Article

Information Theory and Multivariate Techniques for Analyzing DNA Sequence Data: An Example from Tomato Genes

Bal K. Joshi[1] and Dilip R. Panthee[1]

North Carolina State University, Raleigh, North Carolina- 27695, USA
Email: joshibalak@yahoo.com

Abstract

DNA and amino acid sequences are alphabetic symbols having no underlying metric. Use of information theory is one of the solutions for sequence metric problems. The reflection of DNA sequence complexity in phenotype stability might be useful for crop improvement. Shannon-Weaver index (Shannon Entropy, H') and mutual information (MI) index were estimated from DNA sequences of 22 genes, consisted of two gene families of tomato, namely disease resistance and fruit quality. Main objective was use of information theory and multivariate techniques to understand diversity among genes and relate the sequences complexity with phenotypes. The normalized H' value ranged from 0.429 to 0.461. The highest diversity was observed in the gene Crtr-B (beta carotene hydroxylase). Two principal components which accounted for 36.65% variation placed these genes into four groups. Groupings of these genes by both PCA and cluster analysis showed clearly the similarity at phenotypes levels within cluster. Sequences similarity among genes was observed within a family. Diversity assessment of genes applying information theory should link to understand the sequences complexity with respect to gene stability for example stability of resistance gene.

Keywords: *diversity analysis, DNA sequences, principal component analysis, tomato genes*

Introduction

Sequencing of genomic DNA has been started in many organisms in the world and most of the sequences are publically available. International Solanaceae Genome Project (SOL) has started sequencing the genome of tomato. Ten countries namely Korea, China, United Kingdom, India, The Netherlands, France, Japan, Spain, Italy and the United States are involved in the genome sequencing project of tomato (Mueller et al 2005). Tomato is the model plant for the study of a number of economically important traits including fruit development and plant defense (Li et al., 2001; Tanksley, 2004). Major problem with these sequence data is metric problem i.e. difficult to analyze statistically to extract the biologically meaningful information. Use of information theory is one of the ways for handling with sequence metric problem data. Information theory which is based on probability and statistics quantifies information in the categorical form of data e.g. alphabetic sequences of DNA. Entropy is the key measure of information which quantifies the uncertainty involved when encountering a random variable. Another element of information theory is mutual information which is the amount of information in common between two random variables (Schneider, 2003; Atchley et al., 2000).

Many resistant varieties of tomato have been developed through the introgression of resistant genes either from cultivated or from wild species of tomato. More than 8 years is necessary to develop resistant variety through conventional breeding system. However, due to the high mutation rate in pathogen, resistant gene may not be effective for a longer time because of the evolution of mutant pathogen. More stable resistant gene, if possible to identify at early stage may contribute greatly in crop improvement. Similarly, there is a high variation for fruit quality traits of tomato. Many of these genes have been sequenced and data are available in the Solanaceae Genomics Network (SGN) (http://sgn.cornell.edu) (Mueller et al., 2005a; Mueller et al., 2005b). DNA sequences data are added rapidly to the SGN. However, utilization of these data is restricted due to lack of data analysis tools. Using the information theory, various indices and variability can be estimated from the amino acid sequences data and. subsequently analyzed using multivariate techniques (Atchley et al., 1999; Atchley et al., 2000; Atchley and Zhao, 2007; Atchley et al., 2005). DNA sequences are also alphabetic symbols that need to transform these symbols to biologically meaningful variables. Approaches used by Atchley et al. (1999, 2000, 2005) can be used to summarize the DNA sequences data. Multivariate technique, for example Principal Component Analysis (PCA) and cluster analysis would help greatly to draw inference from these DNA sequences data. The objective of this study was to measure the diversity and to study the relationship between sequence variation and phenotype among two gene families namely disease resistance and fruit quality.

Multivariate analysis allows using all available variable information simultaneously producing a single parameter. It has been used for both qualitative and quantitative characters to measure genetic relationships within cereal species e.g. barley (Cross, 1992; Hussaini et al., 1977) and rice (Kanwal et al., 1983). The information generated can be useful for identifying groups of accessions that have desirable characters for further study and for investigating some aspects of crop evolution (Brown, 1991; Cowen and Frey, 1987; Perry and McIntosh, 1991). Among the different methods of multivariate analysis, PCA and cluster analysis are commonly used. PCA is a technique for analyzing relationships among several quantitative variables measured on a number of objects (Ringnér, 2008). It provides information about the relative importance of each variable in characterizing the objects. Cluster analysis can be used to group units according to the similarity for certain characteristics.

Among the different measures of diversity, Shannon-Weaver index (Shannon entropy, H') is being used for qualitative traits (Holcomb et al., 1977; Niwranski et al., 2002; Tolbert et al., 1979) and amino acid sequence data (Atchley et al., 2000, 2005). Principally H' includes both species number and evenness, where a greater number of species increase diversity, as does a more equitable distribution of individuals among species. As species richness and evenness increase, so does diversity. Diversity index provides important information about rarity and commonness of species in a community. The ability to quantify diversity in this way is an important tool for biologists. Alphabetic sequence data can be summarized more precisely with this idea and multivariate techniques. This paper describes the gene diversity in tomato using information theory and multivariate techniques.

Materials and Methods

Tomato Genetic Resource Center (TGRC) (http://tgrc.ucdavis.edu) has listed a total of 1239 genes and their symbols along with their phenotypes. Among them, there are 68 resistance genes related to diseases and 41 genes related to fruit quality of tomato. A total of 22 genes (Table 1) consisting of 8 disease resistance genes, 12 fruit quality related genes and 2 genes from potato genome were used in this study. DNA sequences along with other traits of these genes were downloaded from Solanaceae Genomics Network (http://sgn.cornell.edu).

Multiple alignments of DNA sequences were done in ClustelX2 (http://www.clustal.org/). ClustalX2 calculates the best match for the selected sequences based on the homology concept, and lines them up so that the identities, similarities and differences can be seen. Shannon Entropy (H’) was estimated for each gene, not the site of nucleotides using the FastaEntropy software. FastaEntropy was originally developed for estimating H’ and MI of amino acid sites (column wise) based on the 20 amino acids (see Atchley et al 1999, Atchley et al 2000, (Butte and Kohane, 2000) for details of H’ and MI). We used this software in DNA sequences data. To verify the use of FastaEntropy in DNA sequences data, we estimated H’ of translated amino acids and found the strong association with DNA sequence-based estimates. Normalized entropy value was used to develop the entropy profile of each gene. Normalized mutual information (MI) matrix among the genes was used to generate the scatter biplot considering Principal Components I and II using NTSYSpc (http://www.exetersoftware.com/index.html). Cluster analysis was also carried out based on this MI in NTSYSpc. PCA permits reduction of the complexity or dimension of the problem (Johnson and Wichern, 1988; Ringnér, 2008). The technique consists of reducing the structure of the data matrix starting from a linear model of getting new variables, referred to as principal components. Those principal components with eigen values ≥ 1.0 were selected. Cluster analysis allows one to identify groups of objects or variables that are similar among themselves (Sneath and Sokal, 1973).

Table 1: List of genes and phenotypes along with their location related to diseases resistance and fruit quality of tomato (**http://tgrc.ucdavis.edu/Data/Acc/Genes.aspx**) used in this study

SN	Gene	Allele	Locus Name	Chromosome-arm	Phenotype	Seq, n
Disease resistance genes						
1.	*Asc*	--	Alternaria stem canker resistance	3-long	Resistance to Alternaria stem canker	1496
2.	*Cf-9*	--	Cladosporium fulvum resistance	1-short	Resistance to specific races of *Cladosporium fulvum*	2906
3.	*Hero*	--	Heterodera	4-short	Resistance to potato cyst nematode	4280

SN	Gene	Allele	Locus Name	Chromosome-arm	Phenotype	Seq, n
			rostochiensis resistance		(*Globodera rostochiensis*)	
4.	*Pto*	*1*	Pseudomonas syringae pv tomato resistance	5-long	Resistance to *Pseudomonas syringae* pv. tomato, race zero sensitive to insecticide Fenthion.	2466
5.	*Pto*	*2*	Pseudomonas syringae pv tomato resistance	5-long	Resistance to *Pseudomonas syringae* pv. tomato	2466
6.	*Pto*	*h*	Pseudomonas syringae pv tomato resistance	5-long	Resistance to *Pseudomonas syringae* without Fenthion sensitivity	2466
7.	*Pto*	*Pto-2*	Pseudomonas syringae pv tomato resistance	5-long	Resistance to *Pseudomonas syringae* pv. tomato	2466
8.	*Mi-1.2*		Leucine zipper, nucleotide binding, leucine-rich repeat	6-short	R gene that confers resistance against some species of root knot nematode, and specific isolates of potato aphid, and white fly	3987
Fruit quality genes						
9.	*B*	*1*	Beta-carotene	6-long	Flesh of ripe fruit orange, due to high B-carotene, low lycopene concentrations	1772
10.	*B*	*c*	Beta-carotene	6-long	Increased fruit lycopene content; phenotype similar to *B-og*	1772
11.	*B*	*m*	Beta-carotene	6-long	High B-carotene, low lycopene in ripe fruit	1772
12.	*B*	*og*	Beta-carotene	6-long	Corolla tawny orange; increased fruit lycopene	1772
13.	*CrtR-b*	*wf*	Beta carotene hydroxylase	3-short	White flower	1340
14.	*nr*	--	Never ripe	9-long	Fruits turn color at normal time, but develop full pigmentation slowly and never assume as deep a color as normal	3240
15.	*Psy 1*	*(1s)*	Phytoene synthase 1	3-short	Yellow color of ripe fruit flesh	942
16.	*Psy 1*	*(2s)*	Phytoene synthase 1	3-short	Yellow fruit flesh; lighter yellow flowers	1730
17.	*Psy 1*	*prov4*	Phytoene synthase 1	3-short	Yellow color of ripe fruit flesh	1939
18.	*Psy 1*	*prov5*	Phytoene synthase 1	3-short	Yellow color of ripe fruit flesh	213
19.	*Psy 1*	*y*	Phytoene synthase 1	3-short	Likely allele of r with reddish flesh tones in ripe fruit	303
20.	*rin*	*1*	Ripening inhibitor	5-long	Fruits green at maturity, later turning bright yellow, retarded ripening	1192
Potato genes						
21.	*Star*		Solanum tubersum ankyrin repeat		Putatively involved in quantitative resistance to *Phytophthora infestans*	1508
22.	*RanGA P2*		GTPase-activating protein		A GTPase-activating protein that interacts with Rx, the potato virus X resistance gene.	333

SN= Serial number; Seq= Sequence; n=Number of nucleotide bases

Results and Discussion

Generally mapping populations in tomato is generated by crossing cultivated species with wild relatives which results in maximizing the genetic variation particularly for disease resistance and fruit quality traits (Frary et al., 2000; Fridman et al., 2000) . Therefore, we used these two gene families (disease resistance and fruit quality) to assess and characterize the sequence variation. All available genes with DNA sequences were included in our study.

Tomato is a diploid with 2n= 24 chromosomes and self pollinated species. The eight disease resistance genes are located in chromosomes 1, 3, 4, 5 and 6, mostly in long arm (Table 1). One locus, *Pseudomonas syringe* pv tomato resistance has 4 alleles. DNA sequences of these genes ranged from 1496 to 3987 base pairs. The number of sequences of alleles of *Pto* were the same. With regards to 12 fruit quality related genes, mostly they are located in long arm of chromosomes 3 and 6. Beta carotene gene has 4 alleles with the same number of DNA sequences and *Py-1* gene has 5 alleles with different number of sequences. The allele *prov5* of *Psy-1* has the shortest sequences and never ripe, *nr* has the longest DNA sequences. Two disease resistance genes from potato genome were of 1508 and 333 number of base pairs.

Shannon entropy profiles along with their index (H') of each gene are given in Figure 1. The normalized H' value ranged from 0.4291 to 0.461. All the alleles of beta carotene and *Pto* have the same H' values. Both genes of potato indicated more gene diversity and complexity as compared to tomato genes. Within tomato genes, the highest diversity was observed in the gene, *CrtR-b* (Beta carotene hydroxylase) and the lowest diversity in *Phytoene synthase-1* (*prov5*). Gene of white flower was the most variable and of yellow flower was the least variable. The second highest complexity in DNA sequences was observed in the gene resistance to potato cyst nematode. Entropy values of all the proteins, generated after translating these DNA sequences were higher than their genes, but strong correlation between directly estimated H' values of genes and H' of their translated protein indicates that FastaEntropy can be used to estimate H' from DNA sequences (data not shown). On an average, H' value of disease resistance gene family was higher than fruit quality gene family.

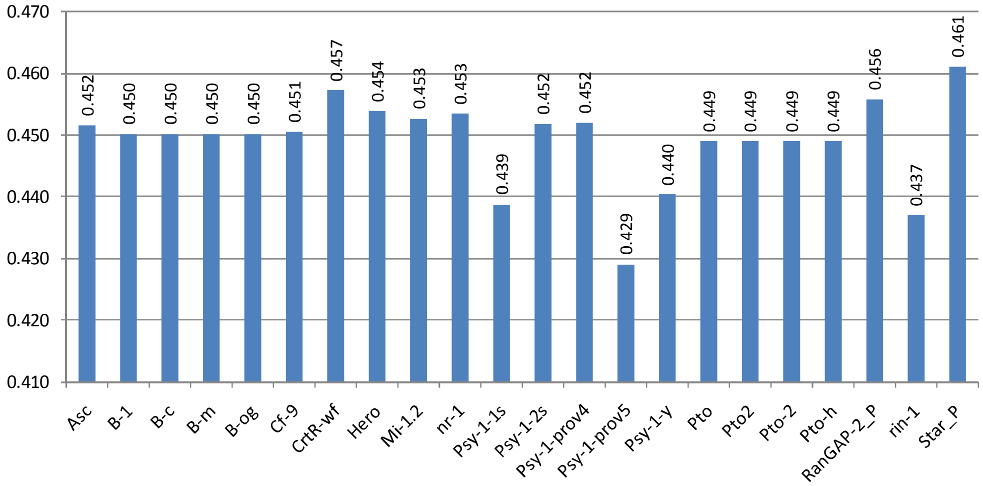

Figure 1. Normalized Shannon entropy profile of tomato genes

The higher the H' value, the more sequence complexity in the gene. Based on this hypothesis the gene with high H' value would be more stable or such gene need long time to take place the evolutionary changes. This is very useful particularly to develop disease resistance variety. Breeder can identify the more stable resistance genes looking on the Shannon entropy profile.

Mutual information values for genes are given in Table 2. The value with 1, for example, among the alleles of *Pto* (*Pseudomonas syringae* pv tomato resistance) and *B* (beta carotene) indicates that the alleles of these genes have the same sequences. The associations of *Psy-1-prov5* (yellow color of ripe fruit flesh) with *Psy-1-2s* and *Psy-1-prov4* were the highest. Association of RanGAP-2 was lower with most of the genes. Mutual information values describe the extent of association between gene pair and zero value between them means they are independent (Atchley et al 1999, 2000). We found all genes having some degree of association with each others.

Result of principal component analysis of sequences of DNA is given in Figure 2. The first seven principal components with ≥ 1 eigen value accounted for 70.03% of the total variance among the genes based on the DNA sequences. The first principal component explained 18.61% and second component accounted for 18.03% variance. Plotting of these accessions along first and second principal components indicated that these genes were placed in four groups. Most of the individual genes make a separate cluster (Figure 3). Clustering these genes can be helpful to identify the similarity among the genes.

Both PCA and cluster analysis separated these genes into four groups. These grouping were also similar phenotypically. For example, genes related to beta carotene make a single cluster and genes related to fruit characters tended to appear together. This indicates that the nature of variation in DNA sequences reflect the similar variation at phenotypic levels.

Table 2. Mutual information matrix among 22 genes of tomato and potato estimated from DNA sequences

Gene	Pto	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Pto2	1	1																			
Pto-2	1	1	1																		
Pto-h	1	1	1	1																	
Psy-1-2s	0.027	0.027	0.027	0.027	1																
Psy-1-prov5	0.081	0.081	0.081	0.081	0.908	1															
Psy-1-prov4	0.025	0.025	0.025	0.025	0.267	0.908	1														
Psy-1-y	0.056	0.056	0.056	0.056	0.637	0.756	0.65	1													
Hero	0.009	0.009	0.009	0.009	0.013	0.031	0.015	0.016	1												
Mi-1.2	0.01	0.01	0.01	0.01	0.01	0.017	0.014	0.013	0.23	1											
B-1	0.011	0.011	0.011	0.011	0.012	0.034	0.011	0.012	0.077	0.064	1										
B-c	0.011	0.011	0.011	0.011	0.012	0.034	0.011	0.012	0.077	0.064	1	1									
B-m	0.011	0.011	0.011	0.011	0.012	0.034	0.011	0.012	0.077	0.064	1	1	1								
B-og	0.011	0.011	0.011	0.011	0.012	0.034	0.011	0.012	0.077	0.064	1	1	1	1							
Psy-1-1s	0.007	0.007	0.007	0.007	0.022	0.021	0.023	0.024	0.009	0.01	0.005	0.005	0.005	0.005	1						
nr-1	0.006	0.006	0.006	0.006	0.01	0.036	0.012	0.009	0.005	0.003	0.005	0.005	0.005	0.005	0.011	1					
rin-1	0.012	0.012	0.012	0.012	0.012	0.029	0.014	0.019	0.006	0.008	0.005	0.005	0.005	0.005	0.007	0.01	1				
Star	0.006	0.006	0.006	0.006	0.018	0.021	0.018	0.054	0.005	0.005	0.014	0.014	0.014	0.014	0.027	0.01	0.01	1			
Asc	0.004	0.004	0.004	0.004	0.003	0.022	0.005	0.02	0.006	0.005	0.008	0.008	0.008	0.008	0.004	0.002	0.011	0.018	1		
CrtR-wf	0.004	0.004	0.004	0.004	0.008	0.021	0.013	0.012	0.007	0.008	0.008	0.008	0.008	0.008	0.009	0.008	0.005	0.02	0.004	1	
RanGAP-2	0.003	0.003	0.003	0.003	0.007	0.028	0.007	0.022	0.006	0.006	0.003	0.003	0.003	0.003	0.008	0.004	0.003	0.014	0.003	0.007	1
Cf-9	0.005	0.005	0.005	0.005	0.007	0.028	0.008	0.023	0.004	0.005	0.008	0.008	0.008	0.008	0.007	0.004	0.004	0.032	0.006	0.007	0.003

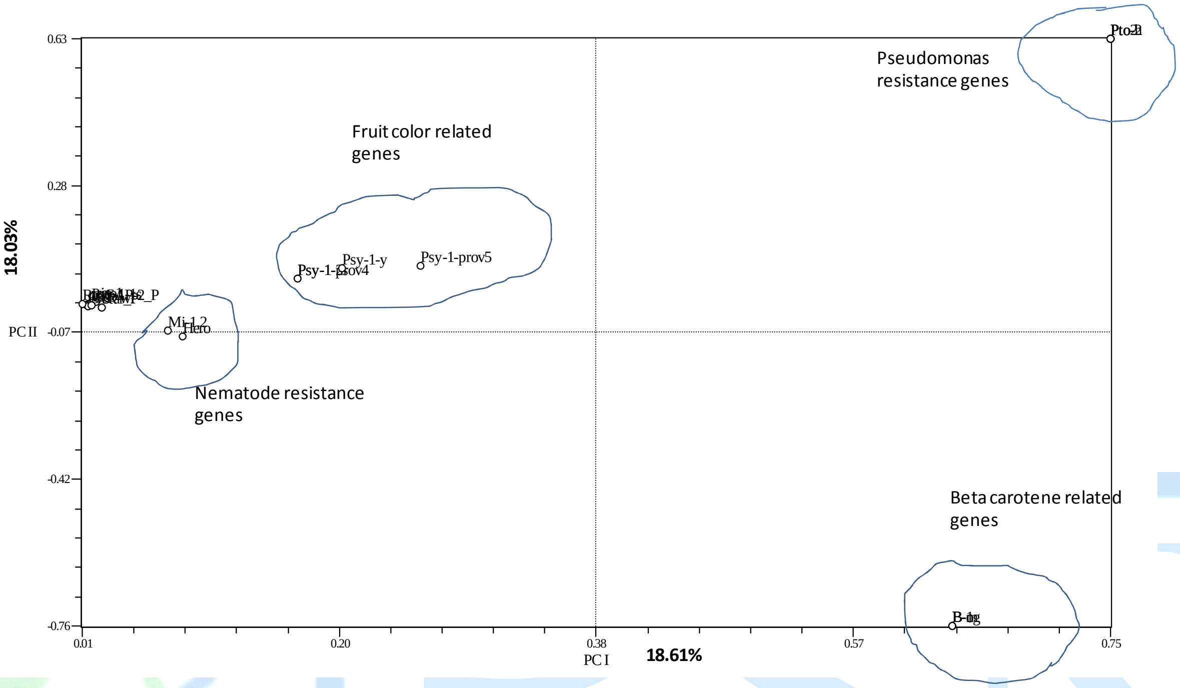

Figure 2. Plotting of tomato genes considering Principal Component I (18.61%) and Principal Component II (18.03%) based on mutual information index calculated from the DNA sequences

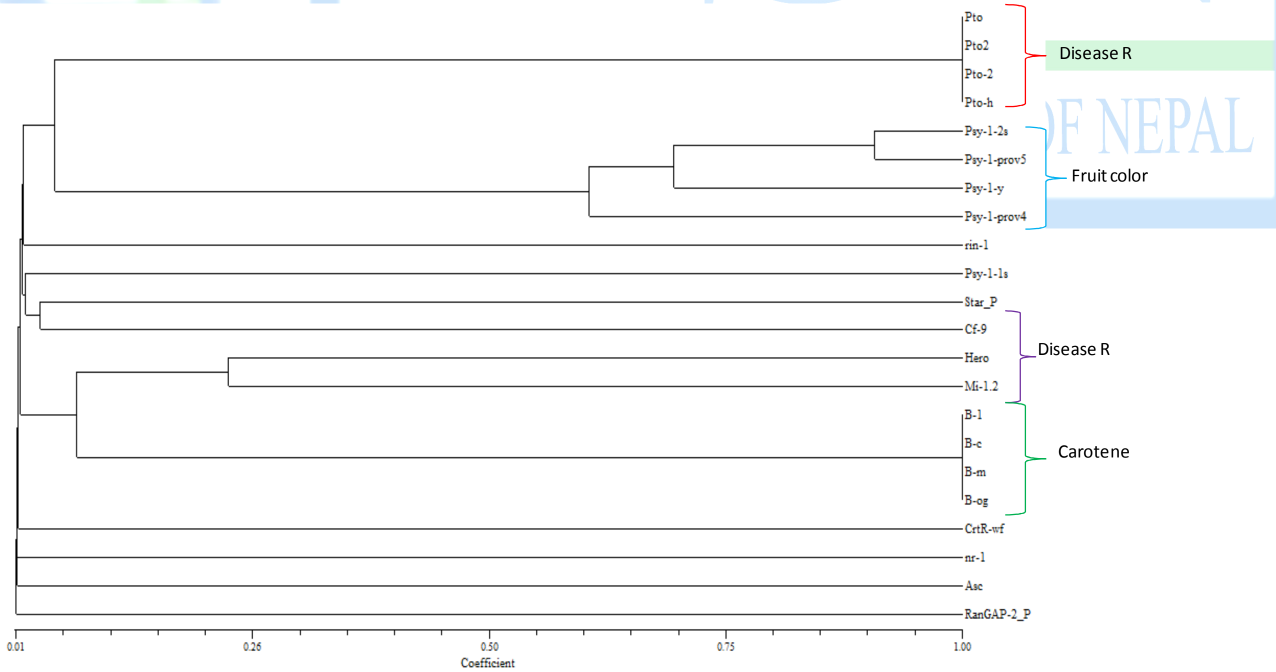

Figure 3. UPGMA cluster analysis of tomato genes generated from the mutual information matrix

Genes within cluster were phenotypically similar. However, one group was consisted of both disease resistance and fruit quality genes. This might be due to role of secondary metabolites in fruit quality and disease resistance. For example tomato fruit contains antioxidant, mainly pigment which is also necessary in defense mechanism. Potato gene, *RanGAP2* made an individual cluster. *Phytophthora infestans* resistance gene from potato clustered with *Cladosporium fulvum* resistance gene of tomato.

Sequences of DNA are deposited in the website by large amount and such data has a sequence metric problem. This problem is now handled following the method of estimating variability and covariability, mutual information index and applying dimension reduction approaches as used in amino acids sequences (Atchley et al 2000, 2005; Atchley and Zhao 2006). Atchley et al (2000) summarized the amino acids sites of bHLH protein by estimating entropy and mutual information values. Information theory is useful to look for pattern in DNA and protein sequences (Schneider, 1999). Information theory has been applied to the analysis of DNA and protein sequences for analyzing sequence complexity from the Shannon-Weaver indices and comparing homologous sites in a set of aligned sequences by means of their information content.

Strong relationship was detected in the present study between DNA variation and phenotype. A large amount of genetic and molecular information has been deposited at http://sgn.cornell.edu (Jiménez-Gómez and Maloof, 2009) that might be useful to explore possible relationships using this technique. Relationship among the diversity index, sequence complexity and phenotype stability of the gene should be considered in the future work to enhance the crop improvement efforts.

References

1. Atchley, W.R., W. Terhalle, and A. Dress. 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. Journal of Molecular Evolution. 48: 501-516.
2. Atchley, W.R., K.R. Wollenberg, W.M. Fitch, W. Terhalle, and A.W. Dress. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Molecular Biology and Evolution. 17: 164-178.
3. Atchley, W.R.and J. Zhao. 2007. Molecular architecture of the DNA-binding region and its relationship to classification of basic helix-loop-helix proteins. Molecular Biology and Evolution. 24: 192-202.
4. Atchley, W.R., J. Zhao, A.D. Fernandes, and T. Drüke. 2005. Solving the protein sequence metric problem. Proceedings of the National Academy of Sciences. 102: 6395-6400.
5. Brown, J.S. 1991. Principal component and cluster analysis of cotton cultivar variability across the US cotton belt. Crop Science. 31: 915-922.
6. Butte, A.J.and I.S. Kohane. 2000. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. Pacific Symposium on Biocomputing. p. 418-429.
7. Cowen, N.M.and K.J. Frey. 1987. Relationships between three measures of genetic distance and breeding behaviour in oats (Avena sativa L.). Genome. 29: 97-106.

8. Cross, R.J. 1992. A proposed revision of the IBPGR barley descriptor list. Theoretical and Applied Genetics. 84: 501-507.
9. Frary, A., T.C. Nesbitt, S. Grandillo, E. Knaap, B. Cong, J. Liu, J. Meller, R. Elber, and K.B. Alpert. 2000. fw2. 2: a quantitative trait locus key to the evolution of tomato fruit size. Science. 289: 85-88.
10. Fridman, E., T. Pleban, and D. Zamir. 2000. A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proceedings of the National Academy of Sciences. 97: 4718-4723.
11. Holcomb, J., D.M. Tolbert, and S.K. Jain. 1977. A diversity analysis of genetic resources in rice. Euphytica. 26: 441-450.
12. Hussaini, S.H., M.M. Goodman, and D.H. Timothy. 1977. Multivariate analysis and the geographical distribution of the world collection of finger millet. Crop Science. 17: 257-263.
13. Jiménez-Gómez, J.M.and J.N. Maloof. 2009. Sequence diversity in three tomato species: SNPs, markers, and molecular evolution. BMC Plant Biology. 9: 85.
14. Johnson, R.A.and D.W. Wichern. 1988. Applied multivariate statistical analysis. 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
15. Kanwal, K.S., R.M. Singh, and J. Singh. 1983. Divergent gene pools in rice improvement. Theoretical and Applied Genetics. 65: 263-267.
16. Li, L., C. Li, and G.A. Howe. 2001. Genetic analysis of wound signaling in tomato. Evidence for a dual role of jasmonic acid in defense and female fertility. Plant Physiology. 127: 1414-1417.
17. Mueller, L.A., T.H. Solow, N. Taylor, B. Skwarecki, R. Buels, J. Binns, C. Lin, M.H. Wright, R. Ahrens, and Y. Wang. 2005a. The SOL genomics network. A comparative resource for Solanaceae biology and beyond. Plant Physiology. 138: 1310-1317.
18. Mueller, L.A., S.D. Tanksley, J.J. Giovannoni, J. Van Eck, S. Stack, D. Choi, B.D. Kim, M. Chen, Z. Cheng, and C. Li. 2005b. The tomato sequencing project, the first cornerstone of the International Solanaceae Project (SOL). Comparative and Functional Genomics. 6.
19. Niwranski, K., P.G. Kevan, and A. Fjellberg. 2002. Effects of vehicle disturbance and soil compaction on Arctic collembolan abundance and diversity on Igloolik Island, Nunavut, Canada. European Journal of Soil Biology. 38: 193-196.
20. Perry, M.C.and M.S. McIntosh. 1991. Geographical patterns of variation in the USDA soybean germplasm collection: I. Morphological traits. Crop Science. 31: 1350-1355.
21. Ringnér, M. 2008. What is principal component analysis? Nature Biotechnology. 26: 303.
22. Schneider, T.D. 1999. Information Theory Primer. Web Document: 1-9.
23. Sneath, P.H.A.and R.R. Sokal. 1973. Numerical taxonomy. Springer.
24. Tanksley, S.D. 2004. The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. The Plant Cell Online. 16: S181-189.
25. Tolbert, D.M., C.O. Qualset, S.K. Jain, and J.C. Craddock. 1979. A diversity analysis of a world collection of barley. Crop Science. 19: 789-794.