Clustering Algorithms for Genetic Analysis with GeneMarker®

February 2014

Tamela Serensits, Wan Ning, Haiguo He, Jonathan Liu, Ph.D.

Introduction

Biological applications of data clustering include phylogeny analysis and community comparisons in ecology, gene expression pattern, enzymatic pathway mapping, and functional gene family classification in the bioinformatics field.[1] It has also been successfully paired with the AFLP analysis technique for a variety of applications.[2]

Amplified Fragment Length Polymorphisms (AFLP®) is a polymerase chain reaction (PCR)-based genetic fingerprinting technique developed in the early 1990's by Keygene*. AFLP technology has the capability to detect polymorphisms in different genomic regions simultaneously. It is also highly sensitive and reproducible. As a result, AFLP has become widely used for the identification of genetic variation in strains or closely related species of plants, fungi, animals, and bacteria. AFLP technology has also been used in criminal and paternity tests, population genetics and linkage studies.[3]

As the results of AFLP are obtained, some researchers turn to novel statistical tools for analyzing the data. An example of this was in 2005 when Fearnley et al. applied a clustering algorithm to AFLP data.[4] The results helped differentiate the relationship between closely related strains of Yersinia enterocolitica, a bacterium that infects several species including humans, pigs, sheep, and cattle. The study found clustering analysis of AFLP data to be highly discriminatory.

GeneMarker is an easy-to-use, accurate fragment analysis tool and can perform analysis on up to 1,000 lanes of four or five color data sets generated by either slab gel or capillary electrophoresis. It is a unique genotyping tool as it is compatible with files from all major capillary and slab gel electrophoresis systems including ABI files (*.FSA, .*AB1, *.ABI), SCF files, MegaBace files (*.RSD, *.ESD), SpectruMedix files (*.SMD, *.SMR), and Beckman files. **AFLP is a registered trademark of KeyGene, N.V.*

Procedure

There are two types of data clustering: hierarchical and partitional. Partitional clustering includes the K-means and Self-Organizing Map methods. Hierarchical clustering is the second method of clustering and is the method that is implemented in GeneMarker. Hierarchical Clustering treats each data point as a single cluster and successively merges clusters until all points have been merged into a single remaining cluster. Hierarchical clustering is often represented as a dendrogram. In GeneMarker, the hierarchical algorithm is agglomerative and establishes clusters from the bottom up.

Distance Measure

The first step in hierarchical clustering is to select a distance measure. GeneMarker distance options include Euclidean Distance, Correlation Coefficient, and Percentage of Same Genotypes. Euclidean Distance is the straight line distance between two points in two or three dimensional space. The equation is essentially the same as that for determining the length of the hypotenuse of a triangle – computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. We have simplified this equation (below) in GeneMarker. The Correlation Coefficient is based on the Pearson Correlation equation and is a statistical concept that quantifies the level of relationship between two sets of measurements. It is a measure of similarity where two values that are perfectly correlated have a distance of 1.00. Percentage of Same Genotypes is simply the number of similar genotypes divided by the total number of genotypes.

The following are GeneMarker's clustering algorithms:

Euclidean Distance: $\sum_{i=1}^{n} |(x_i - y_i)|$ **Correlation Coefficient:** $r \equiv \sqrt{bb'} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$

Linkage

In addition to a distance measure, the type of linkage needs to be applied. GeneMarker has three options: Single, Complete, and Average linkage. Single linkage measures the minimum distance between two clusters. Clustering using single linkage tends to produce an effect called chaining where single genes are added to clusters one at a time. Complete linkage is the opposite of single linkage. It measures the distance between the farthest two points in the clusters. Complete linkage performs well when the clusters are well defined with minimal noise. Average linkage defines the distance between two clusters as the mean distance between all points in the clusters. It is important to note that choosing different linkage measures results in different cluster diagrams. We demonstrate in the Results section how different distance measure and linkage analysis settings have an effect on how the data are analyzed.

Results

Notice how when just the distance measure is changed (Fig 1 & 2), the basic overall structure is similar, however; on closer examination the fine structure of ordering within the main clusters differs. The samples with "3" as the first character in the file name are grouped, as are the samples with the number "4".

The sole "7" sample is grouped in its own cluster in both examples. These results are as expected.

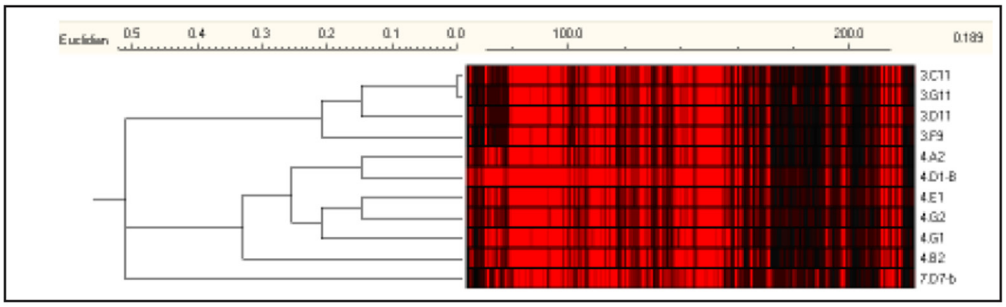

Fig. 1 Euclidean Distance Single Linkage

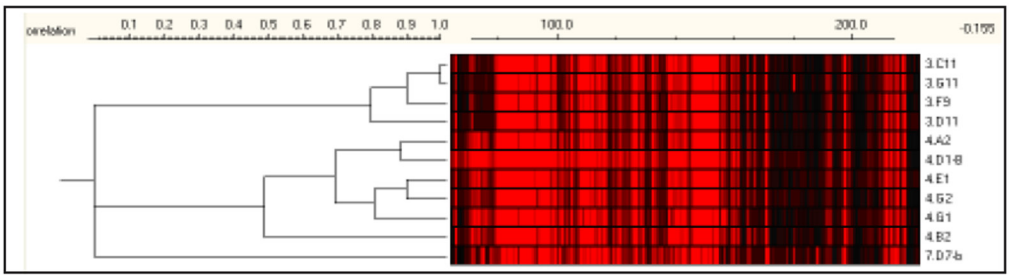

Fig. 2 Correlation Coefficient Single Linkage

When altering the analysis based just on linkage type and holding the distance measure constant (Figs 2-4), we see that the overall structure remains the same, however; the finer structure is greatly affected. Notice how in single linkage (Fig 2) the three main groups are independent of one another; where in complete linkage (Fig 3), the "4.B2" sample is separated from the main group. This is representative of complete linkage's tendency to form smaller, compact clusters. It can also be seen from this example how average linkage (Fig 4) is an amalgam of single and complete linkage.

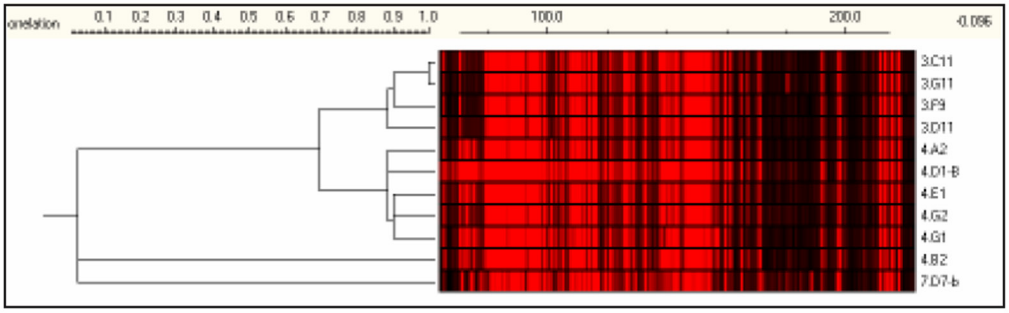

Fig. 3 Correlation Coefficient Complete Linkage

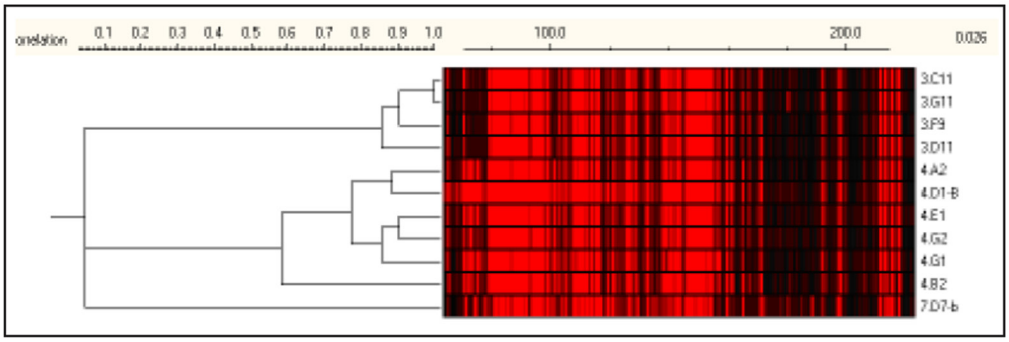

Fig. 4 Correlation Coefficient Average Linkage

In addition to dendrograms, GeneMarker outputs a Matrix Report to save as a Text (.txt) file. As mentioned in the Introduction, when the Correlation Coefficient distance measure is applied to the data, a value of 1.00 indicates a perfectly correlated pair. This can be observed in the Matrix Report where the row and column of the same sample meet (Fig 5).

	1	2	3	4	5	6	7	8	9	10	11
1	1.000	1.000	0.902	0.884	0.636	0.551	0.487	0.551	0.487	0.327	0.393
2	1.000	1.000	0.902	0.884	0.636	0.551	0.487	0.551	0.487	0.327	0.393
3	0.902	0.902	1.000	0.797	0.568	0.487	0.617	0.694	0.617	0.280	0.333
4	0.884	0.884	0.797	1.000	0.452	0.388	0.339	0.388	0.339	0.388	0.265
5	0.636	0.636	0.568	0.452	1.000	0.884	0.797	0.884	0.797	0.636	0.062
6	0.551	0.551	0.487	0.388	0.884	1.000	0.694	0.776	0.694	0.551	0.024
7	0.487	0.487	0.617	0.339	0.797	0.694	1.000	0.902	0.808	0.694	-0.007
8	0.551	0.551	0.694	0.388	0.884	0.776	0.902	1.000	0.902	0.551	0.024
9	0.487	0.487	0.617	0.339	0.797	0.694	0.808	0.902	1.000	0.487	0.163
10	0.327	0.327	0.280	0.388	0.636	0.551	0.694	0.551	0.487	1.000	0.024
11	0.393	0.393	0.333	0.265	0.062	0.024	-0.007	0.024	0.163	0.024	1.000

Fig. 5 Clustering Report

Discussion

As we have seen, the linkage method and distance metric chosen produce different clustering results. The following tables demonstrate the strengths and weaknesses of each parameter.[5]

	Euclidean Distance	*Correlation Coefficient*
Strengths	Geometric interpretation	Powerful
	Retains up/down regulation scaling	Detects positive and negative correlations
	Detects magnitude of changes without scaling	Scale invariant on centered data
Weaknesses	Results depend on scaling used	Assumes linearity
	Cannot detect negative correlations	Susceptible to outliers

	Single Linkage	*Complete Linkage*	*Average Linkage*
Strengths	Simple analysis	Highly informative	Most commonly used
	Useful when data clusters well defined but have an irregular shape	Produces small, compact, well-defined clusters	Middle-road between the extremes of single and complete linkage
Weaknesses	Chaining effect - single clusters added one at a time	Does not perform well on noisy data	Measure is an average, not an actual distance, making analysis more difficult
	Sensitive to outliers	Forms many clusters	

Which parameters you choose are up to you – there is no right answer. We recommend that you apply all distance measures and linkage algorithms to your cluster analysis and look at the results to determine which method is right for your data.

Acknowledgements

We would like to thank Heidi Meudt PhD at the Museum of New Zealand, Te Papa Tongarewa, New Zealand, and Andrew Clarke at Massey University, New Zealand for their collaboration. We would also like to acknowledge Haiou Hu for her contribution in developing GeneMarker's clustering algorithms.

References

1. **Data clustering in life sciences.** Y Zhao, G Karypis. *Molecular biotechnology.* 2005. 31 (55-80).
2. **Almost Forgotten or Latest Practice? AFLP applications, analyses, and advances.** HM Meudt, AC Clarke. *Trends in Plant Science.* (in press).
3. **AFLP: a new technique for DNA fingerprinting.** P Vos, R Hogers, M Bleeker, M Reijans, T Lee, M Hornes, A Frijters, J Pot, J Peleman, M Kuiper, M Zabeau. *Nucleic Acids Research.* 1995. 23 (4407-4414).
4. **Application of Fluorescent Amplified Fragment Length Polymorphism for Comparison of Human and Animal Isolates of Yersinia enterocolitica.** C Fearnley, SLW On, B Kokotovic, G Manning, T Cheasty, DG Newell. *Applied and Environmental Microbiology.* 2005. 71 (4960-4965).
5. **Microarray Bioinformatics.** D Stekel. *Cambridge University Press.* 2003. (139-182).