Inferences about transmission using genetic data: not seeing the forest for the trees (the statistical woodchopper's dilemma)

R.M. Weigel*, B. Qiao. Dept. of Veterinary Pathobiology, University of Illinois, Urbana, IL.

Summary: In studies of pathogen transmission, cluster analysis is typically used to classify isolates according to their genetic relatedness. In this study, the suitability of several cluster analysis algorithms, as well as multidimensional scaling (MDS) with a superimposed minimum spanning tree (MST), are examined for use in making inferences about transmission, using a sample of 36 isolates of Salmonella from 3 swine farms. Complete linkage is recommended for identifying clusters of closely related isolates. The Neighbor Joining method is recommended when identifying the degree of divergence from a common ancestor is desirable. MDS in conjunction with MST may distort distances and thus fail to identify the most closely related isolates.

Introduction: Inferences regarding transmission of pathogenic microorganisms are based on the genetic relatedness among isolates. The inter-relationships among isolates are usually resolved using cluster analysis. The resulting tree structure identifies close genetic relatives, and can be interpreted as representing transmission links. However, there are numerous clustering algorithms; which algorithm is selected may affect the classification of organisms into groups of closely related samples. The resulting dendrogram depicts clonal genetic proliferation, which may not be valid for some genetic material in bacteria, where horizontal transmission is possible. Thus, alternatives to cluster analysis that map cases in multidimensional space may more accurately reflect the processes of gene transmission. Multidimensional scaling (MDS) is an alternative statistical method for representing the relatedness of cases in multidimensional space. Superimposing a minimum spanning tree (MST) on a multidimensional scaling configuration will connect nearest neighbors and provide a basis for identifying transmission links.

Methods: Available for analysis were 36 isolates of *Salmonella enterica* collected from 3 swine farms in Illinois (Barber et al., 20002). Each isolate was genotyped using 3 restriction enzymes, Spe I, Xba I, and Avr II, with the fragments separated using pulsed field gel electrophoresis. The genetic distance between isolates was calculated separately for each enzyme *i* as $d_{i_{xy}} = 1 - \left[2n_{i_{xy}} / \left(n_{i_x} + n_{i_y}\right)\right]$, where n_{xy} is the number of fragments matching in size for samples x and y, n_x and n_y are the number of fragments in sample x and y, respectively. A 5% error tolerance range was allowed for designation of matches. The genetic distance between samples x and y over 3 enzymes was calculated as the Euclidean distance $D_{xy} = \sqrt{\sum_{i=1}^{3} d_{i_{xy}}^2}$. The 3-dimensional genetic distance matrix was the basis for further analysis. Hierarchical cluster analysis was used to represent graphically the genetic relationships among the Salmonella isolates. Several clustering algorithms - single linkage, complete linkage, average linkage [UPGMA] (Everitt et al., 2001), and nearest neighbor joining (Nei & Kumar, 2000) - were compared to examine their effect on classification of isolates into clusters. A cluster was defined as the cases joined at a distance ≤ 0.15 (i.e., ≥

85% of fragments matching in size). Multidimensional scaling (Kruskal & Wish, 1978), using a 3 dimensional solution, was used to plot the isolates in 3D space to represent their proximities in the genetic distance scale. MDS configures cases in multidimensional space such that the difference between the derived distances and the original distances are minimized. A minimum spanning tree [MST] (Xu et al., 2002) was imposed on the MDS configuration to identify nearest genetic neighbors. An MST links cases such that the total branch length of the tree is minimized.

Results: Single linkage cluster analysis resulted in the fewest number of clusters (6) and complete linkage had the most (12). With single linkage, all 12 isolates from farm PG were in one cluster, whereas with complete linkage the 12 isolates were distributed over 3 clusters, and 1 isolate was an outlier, not associated with a cluster. The UPGMA algorithm had intermediate results (8 clusters, 2 for farm PG with no outliers). For the Neighbor Joining method (Fig. 1), 6 clusters were apparent. For all algorithms, clusters were homogeneous for farm of origin. MDS with 3 dimensions had a configuration with STRESS1= 0.14. The closest links are identified by the Minimum Spanning Tree overlaid on the 3D MDS configuration (Fig. 2), which identified 4 large clusters, 2 each for farm RB and 1 each for farms FF and PG.

Discussion: Although each algorithm produces a unique cluster structure, the closest neighbors will be the same for each method. However, if there is any interest in identifying groups of isolates with a shared phylogenetic and transmission history, single linkage, which forms clusters with high diversity, is not recommended. In this case, complete linkage, which only connects closely related isolates, is recommended. Neighbor Joining has an advantage in representing the degree of divergence from a common ancestor, and is the preferred method if this information is useful. MDS does not produce a hierarchical dendrogram, and thus can represent multiple origins of genetic similarity (clonal proliferation and horizontal transmission). Superimposing an MST on a MDS configuration, connecting nearest neighbors, links more directly those isolates that are most closely related to each other, but may distort the relative distances among most closely related isolates. The Neighbor Joining method can be reconfigured as an unrooted tree, which reduces the impression of unilineal clonal proliferation, and also have the same form as an MST without distorting distances between closest neighbors. Among the methods compared, the Neighbor Joining method has numerous advantages, its only drawback being greater difficulty in identification of clusters. This is a minor disadvantage when the focus is on identifying nearest genetic neighbors, which is typical in transmission studies.

References:

1. Barber DA, Bahnson PB, Isaacson RE, Jones CJ, Weigel RM. 2002. Distribution of *Salmonella* in swine production ecosystems. *J Food Prot* 65:1861-1868.
2. Everitt BS, Landau S, Leese M. 2001. Cluster Analysis, 4th ed. Arnold. London.
3. Kruskal JB, Wish M. 1978. Multidimensional Scaling. Quantitative Applications in the Social Sciences, no. 11. Sage Publications. Newbury Park, CA.
4. Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press. New York.

5. Xu Y, Olman V, Xu D. 2002. Clustering gene expression data using a graph-theoretical approach: an application of minimum spanning trees. Bioinformatics 18:536-545.

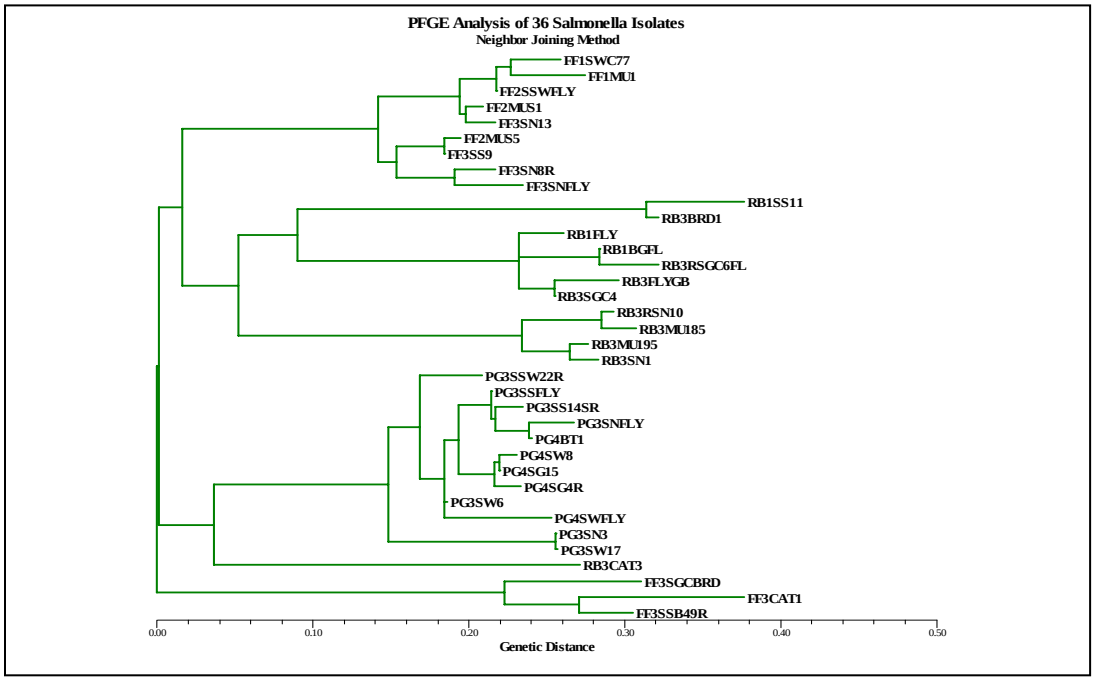

Figure 1. Cluster Structure with Neighbor Joining Method

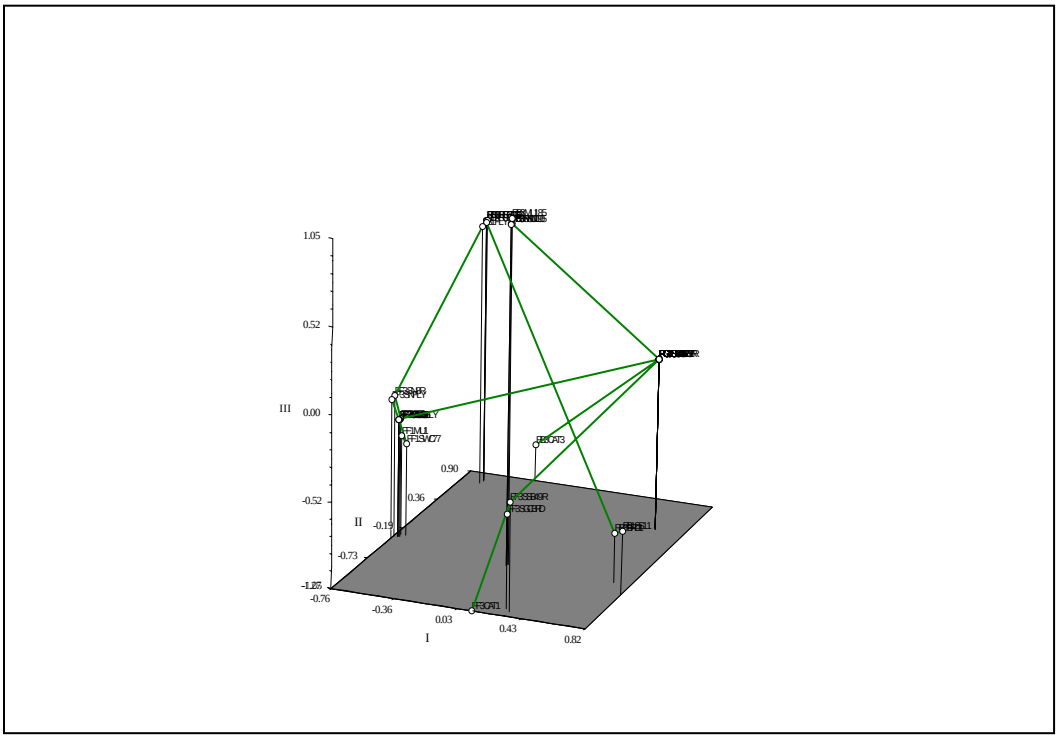

Figure 2. Minimum Spanning Tree superimposed on Multidimensional Scaling configuration