A Tale of Two Methods: Visualizing Pedigree and Population-Level Data in Molecular Ecology

Matt Thorstensen\*, William Hemstrom\*, and Andrea Schreier

Introduction

Visualizing multidimensional data, such as with pedigree or within-population diversity is difficult in molecular ecology. Data visualization can be a helpful tool for first-pass investigations on a data set when beginning analysis and for presenting results in informative figures. For example, STRUCTURE plots (citation) show information about population structure and hybridization in clear ways, but population substructure or hierarchical structure are difficult phenomena to display. Principle components analyses (PCA’s) have been used to show population structure, but with a focus on maximizing distances between clusters, are poor at showing within-population variance. Similarly, PCA’s struggle to display pedigree information in one plot, with family groups usually impossible to show in a 2 or 3-dimensional plot.

We present here a method called t-Distributed Stochastic Neighbor Embedding (t-SNE; Van Der Maaten & Hinton 2008), which is a dimension reduction algorithm that displays population genetic data in 2 or 3 dimensions while making compromises between within- and between-cluster distances so that fine-scale as well as large-scale patterns in data may be visualized. This is a flexible algorithm which we apply to microsatellite and single nucleotide polymorphism (SNP) data sets used to study both population structure and pedigrees.

t-SNE has been used in machine learning and biology for dimension reduction and data visualization to great effect. For instance, t-SNE has been used to support research on deep convolutional networks being trained in visual recognition (Donahue et al. 2014). In biology, t-SNE has been applied to cancer and epileptic seizure research to help identify tumor subpopulations that affect patient outcomes and for detecting epileptic seizures (Abdelmoula et al. 2016; Birjandtalab et al. 2016). In transcriptomics, t-SNE has been useful for differential gene expression, such as in islets of Langerhans within the human pancreas (Muraro et al. 2016). One example of t-SNE applied to SNP data shows that it is more effective than a PCA at resolving human population structure (Platzer 2013).

Briefly, t-SNE works by considering its input data set a high-dimensional matrix that it attempts to display in 2 or 3-dimensions, as a user desires. In the process of reducing dimensions, it considers the distance between data points conditional probabilities, and the similarity between data points is calculated as the conditional probability that one point would pick another as its neighbor randomly following a Gaussian distribution (Van Der Maaten and Hinton 2008). Conditional probabilities are also calculated between points in the low-dimensional space for plotting using a Student-t distribution with 1 degree of freedom (Van Der Maaten and Hinton 2008). This Student-t distribution with 1 degree of freedom has a heavy tail that leads to the compromise of within-cluster and between-cluster distance displayed in a t-SNE plot (Van Der Maaten and Hinton 2008). t-SNE then attempts to minimize the difference between its low and high-dimensional plots by minimizing the sum of Kullback-Leibler divergences over all data points, or in other words attempts to minimize how much one probability distribution differs from the second. A more complete explanation of how the t-SNE works and important parameters to consider when running the algorithm can be found in Van Der Maaten and Hinton (2008).

As raw input data, t-SNE takes data sets where each row is a member of the data set and each column is a variable measured for that member. In transcriptomics, for instance, a row might be an individual organism and a column might be the counts of RNA reads for a particular contig. We realized that a presence-absence format for alleles measured as microsatellites or SNP’s follows this data format, and that t-SNE could be used for questions commonly asked in molecular ecology.

To test the utility of t-SNE for visualizing genetic and genomic data, we plot microsatellite and SNP data sets collected to answer pedigree and population-level questions in several species. For all t-SNE plots, we present a corresponding PCA plot to explore how well each dimension reduction algorithm operates under different conditions. We also develop an R package to easily reformat data, run t-SNE’s and PCA’s, and plot the outputs.