Michael Boodoo

Genomics Outline Draft

Multidimensional Scaling as a Tool for Quantitative Comparison of SCN5A Transcripts

* I am thinking of focusing on comparing the transcripts of disease causing transcripts but I may or may not include a comparison of orthologs to track the evolution of the gene depending on how easy the method is.

Introduction

- **SCN5A**
  - Structure (alpha 5 subunit sodium gated channel)
  - Function (heart physiology)
  - Role in disease (mutations that result in a large class of related diseases)

- **Multidimensional Scaling**
  - What is it? (thorough but easy to read intro on the method and basic mathematical theory on which it is based)
  - Why is it useful? (there are lots of other algorithms in use in bioinformatics for analyzing and comparing genes/protein sequences, why is this particular method useful.
  - Quantitative mapping (unlike phylogenetic trees where the branch lengths are arbitrary, MDS is a quantitative map that compares the similarity of objects to one another using distance.)
  - Dimensions (2-dimensional and 3-dimensional (and beyond) are possible and make more accurate maps, but the increase in dimensions makes interpretation harder since each dimension is it itself a basis of comparison.
  - Basic examples of how this technique is used (widely used in psychology and other scientific fields (for instance, in analyzing the voting patterns of people, or in bioinformatics for visualizing the lineages of groups)
  - Visualization (the key output of MDS is the plot that it produces. The whole goal of MDS is to map objects into a *N*-dimensional space where the *distances* between the objects are conserved as well as possible.)

- **Before running the analysis**

- Define the distance function (embed objects into a dissimilarity matrix that preserves distance into the output. The mathematical theory for this is based on linear algebra which even if some people haven't taken is easy to explain as it is basically a way of solving systems of equations in algebra. I will explain all this during the presentation):

$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix}.$$

Figure 1: Source: wikipedia

- Using R as the program (briefly explain what R is; a programming language that is used for statistics, similar to SPSS or excel)

- Define the problem (what variables and how many do I want to compare? In this case I want to compare the amino acid transcripts of my gene in the native human and disease causing states. I will have plenty of variables given that there are hundreds of known mutations corresponding to human disease phenotypes).

- Obtain the input sequence data (through ensembl and references)

- **Running the Program (This begins the methods section... and will include the code and steps as to how I am performing this analysis in R; all steps and source code will be included so anyone can do what I did step by step)**