Multidimensional Scaling Analysis of Voltage-Gated Ion Channels

Michael Boodoo

5/14/2014

INTRODUCTION

Multidimensional scaling (**MDS**) is a set of data analysis techniques that displays distance data in a geometrical format. It was originally used in psychometrics, where it was used to quantify peoples judgments about the similarity between members of a given set of objects. Unlike a scatter plot, MDS allows visualization of the structure of set of objects from data that is not point like. For example, if you have square matrix (square table with each axis listing a set of cities in the same order) where each entry corresponding to the distance between the given to cities and a grade, this data cannot be easily visualized by simple scatter plot. In MDS, each point is represented by a point in a multidimensional space. This type of analysis translates easily to the analysis of protein sequences in the context of amino acid sequence space, which will be defined later. In MDS, points are placed in a space so that the distances between these points are strongly correlated to the similarities or dissimilarities. Referring to proteins, two proteins with very similar amino acid sequences can be represented as 2 points close together in the amino acid sequence space, whereas 2 dissimilar proteins would be represented by 2 points are far apart. The space and we referring to can be non-Euclidean and can many more dimensions, although MDS can reduce the number of dimensions to create an image that can be visualized and interpreted more easily. There are many types of MDS, such as classical, weighted, and replicated depending on the nature of the data in the type of analysis needed. Classical MDS will be used here because it relies only on one matrix and uses Euclidean distance to model dissimilarity [6]. This distance is defined as [7]:

$$d_{ij} = \sqrt{\sum (x_{ia} - x_{ja})^2}$$

which represents the distance between points i and j on dimension a. This is Euclidean distance between the 2 points represent a straight line distance in real space. When the coordinates of points are known, the distance can be easily calculated using Pythagoras' theorem. In the case of amino acids, we cannot know the coordinates of the sequence in the amino acid sequence space, which will be defined later. The lack of coordinates is also another reason why scatter plots are not applicable for comparing amino acid sequences with one another. Another way of tackling the distance between a pair sequences must be used. Using linear algebra, this distance is defined by [7]:

$$d_{ij} = [(x_i - x_j)(x_i - x_j)^r]^{\frac{1}{2}}$$

where x_i is the *i*th row of X and contains the r coordinates of the *i*th point on all r dimensions. Thus, the distances between points i and j are contained in an -by-n symmetrical matrix. A classical MDS in which the similarities between objects or quantitative (such as in the distances in miles between a set of cities) is also known as a metric MDS. The early types of this kind of MDS required that the data the in terms of dissimilarities, complete and symmetric. It also required that the distances be Euclidean [6]. Metric MDS is actually a superset of MDS and has the benefit in that the distance matrices can be customized with various weights. This corresponds beautifully to the amino acid substitution matrices such as BLOSUM and PAM that weigh amino acid substitutions with various weights depending on how likely the given substitutions were seen in their respective data sets. While multidimensional scaling is powerful method for analyzing these complex data sets, it has unfortunately been ignored in sequence analysis. Multiple sequence alignments lend themselves exceedingly to MDS because it yields a

low dimensional representation of the distance matrix from and aligned set of sequences, whether they be of nucleotides or amino acids. Regardless of the sequence length, this method takes a matrix of Euclidean distances between all pairs of sequence, and finds a space in which these distances are preserved.

In 1992, Higgins [8] discussed the potential use of MDS in analyzing sequence alignments using MDS is a way of analyzing evolutionary distance between sequences. The only hard part is find a measure of the distance between the multiple sequence alignments that is Euclidean. The simplest distance function, which is the square root of the percentage difference is used since it ignores positions in the alignment where there is a gap. In 1971, J. Gower [10] described a general way in which a measure of similarity could be converted into a Euclidean distance. In the case of a multivariable character, such as an amino acid sequence, a coefficient known as Gower's coefficient, is getting by the number of matching characters between 2 objects, divided by the number of total characters. This is a simple ratio I can be converted into a Euclidean distance by subtracting this value from wanted taken square root. Thus, the Euclidean distance between 2 amino acid sequences can be calculated as the square root of: a number of places in the alignment where the two sequences differ (ignoring any gaps), divided by total number of sites. Stated mathematically, for 2 sequences *i* and *j* with *m* identical amino acids from *n* total positions, the Euclidean distance between them is [7]:

$$d_{ij} = [\frac{(n-m)}{n}]^{\frac{1}{2}}$$

In genomics and bioinformatics, the multiple alignment of homologous sequences is a crucial source of information that is used to analyze the evolution and sequence function relationships of protein families. Tree-based or space-based methods are 2 ways that sequences

can be compared, that both rely on the multiple sequence alignment of homologous sequences. The advantage of space-based methods, such as metric multidimensional scaling, is that it does not assume a specific structure to the data. This is in contrast to phylogenetic tree-based method, it assumes that there must be a hierarchical structure upon which phylogenetic relationships can be deduced. MDS only uses a matrix of distances between elements, as defined above, to visualize the given elements in a low dimensional space in which the distances best approximate the original distances. The fact that protein sequences can be visualized in low dimensional space is very important, because the sequence space of a several thousand amino acid protein can't possibly be visualized [5].

The distance in this case refers to the numerical value of the difference of the proteins in **sequence space**. Sequence spaces are well defined mathematical structures consisting of a set of vectors that contained ordered sequences of numbers. It is easy to generalize DNA or amino acid chains as sequences of nucleotides or amino acids to form a vector space that contains an infinite number of proteins, the sequences of which can be anything. Since the set of all naturally occurring proteins makes up a infinitesimally small percentage of all "theoretically" possible sequences, this results in vanishingly small regions of the space populated by the naturally occurring proteins. The naturally occurring set of proteins is further decreased by the fact that many point mutations result in proteins that are structurally similar in terms of folding [1]. This means that a given family of proteins that occupy different regions in sequence space are actually functionally the same protein given that they fold in the same exact way. This is commonly seen in mutations that swap 2 amino acids that have the same chemical behavior in a region of the protein away from the active site. It is expected that a given ortholog will lie near its "cousin" protein resulting in clusters of protein groups. This can't be visualized easily because a

polypeptide sequence space is an *n-dimensional* spaces where n is equal to number of amino acids in a protein. Since many proteins can be thousands of amino acids long, any protein consisting of more than 3 amino acids can't be visualized easily. Even in 3 dimensional sequence space, there are $20^n = 20^3 = 8000$ possible amino acids that can be constructed. In 4 dimensions, which we can't visualize, there are a staggering 160,000 possible proteins. Proteins that are several thousand amino acids long thus exist in a multi-thousand dimension space where the possible number of total protein sequences quickly exceeds the number of particles in the observable universe.

Given a MSA, a distance matrix must be generated before the MDS can be run. 2 major decisions in regards to distance matrices are involved here. A **difference** matrix is based on the physical distances between a pair of points, or in the context of bioinformatics, of a pair of protein sequences. Although abstract, sequence spaces of n-dimensions can be mathematically compared with one another easily which is what MDS does. Referring back to the difference matrix, a more suitable method (in the context of visualizing *ortholog* evolution) for comparing MSA's is the **similarity** or **dissimilarity** matrix. This analysis is more favorable for ortholog analysis because it uses a **substitution** matrix to account for the fact that amino acids are more likely to evolve into certain amino acids more than others. This is important when analyzing protein evolution because given that a certain mutation will either be deleterious, neutral or beneficial, and that the emergence of orthologs by definition relies on mutations that resulted in functional protein in these new species, the rate of substitution has to be taken into account to accurately infer evolutionary relationships. Substitution matrices help evaluate the probability that a given amino acid (or nucleotide) is replaced with another. Early matrices such as the Jukes-Cantor matrix (1969) assumed that mutations occurred with equal probability, leading to

the conclusion that mutation rate was a linear function of time. Later matrices began to discriminate the differences in which amino acids or nucleotides could change. The Kimura matrix (1980) differentiated between the transition and transversion rate of nucleotides. Later matrices such as **BLOSUM** and **PAM** used empirical data to determine the probabilities of residue mutation. PAM is based on a dataset of 1572 mutations in 71 conserved proteins (where the sequence differences in each tree were no more than 15% varied). Thus, this method is useful for analyzing closely related proteins, although mutation probabilities over longer lengths of time can be obtained by multiplying PAM by itself n times, resulting in a PAM-n matrix. PAM tells us that if we have 2 residues **x** & **y**, the calculated probability will be such that a given x will have been replaced by y [3]. BLOSUM substitution matrices help remedy the bias caused by using highly similar proteins. The BLOCKS database of un-gapped sequences was used to create BLOSUM matrices at various similarity levels, denoted by a number. This number corresponds to a similarity threshold at which the given sequences will condense, being counted as *one*. BLOSUM62, a widely used substitution matrix, has a threshold of 62% meaning that if a given set of sequences is greater than or equal to 62% identical, they will be counted as 1 sequence. A very low BLOSUM matrix thus consolidates sequences that are somewhat similar, placing the emphasis on only the most divergent sequences [4].

One of the most useful features of MDS is that it can "project" **supplementary** elements onto the active space. In the context of comparing orthologs, this means that a given protein family of one species can be physically *mapped* onto the active space of interest. In the case of sodium voltage gated channels, the family of all sodium channels of the chimpanzee for instance, can be mapped onto the active space of human sodium channels. This allows comparison of the chimp protein family in the perspective of the human family. If the human set of sodium

channels was made the supplementary set, and then projected onto the chimp protein active space, we would visualize the set of human proteins in the perspective of the chimp family. Large protein families that are clustered into known subgroups can thus be compared with orthologous subgroups to track their evolutionary drift over time. To create the **active** space, we create some sort of distance matrix with the group of genes of interest. Usually, since these genes are closely related, such as a set of human GPCRs, a Euclidean distance matrix is sufficient. It's not necessary to create a distance matrix based on amino acid dissimilarities, otherwise known as a dissimilarity matrix, since we are just creating a two-dimensional and three-dimensional space onto which we can project orthologous genes of interest. As shown by [5], human GPCR sequence spaces obtained different distance matrices do not reveal dramatic differences and it is seen that the overall patterns are maintained conserved. Distance matrices obtained with difference scores or their square roots are very similar and are based solely on a distance *p* that is simply the proportion of the amino acid sites that differ from one another the total number of sites. Thus, the difference score can range from 0, for identical sequences, to 1, for completely different sequences. MDS serves mainly as a tool for the visualization of spatial relationships between elements. The interpretation of the outputs, like in statistics, depends entirely on the type of question you want to answer and is in many ways the hardest part of the analysis.

Pele et al. [5] suggests use of MDS to analyze evolution of orthologous protein sets, and this can be done from the point of view of any MSA depending on what you make the active space. The active space is thus the "canvas" that other orthologs get painted onto. Voltage-gated ion channels in humans are part of a paralogous family of genes that code the proteins responsible for signal transduction. The 3 main groups of voltage-gated ion channels are sodium, potassium and calcium. An isolated sodium channel alpha subunit of interest, *SCN5A* is

implicated in a wide variety of human cardiac diseases such as Brugada syndrome, Long QT-syndrome, atrial standstill, idiopathic ventricular fibrillation and primary cardiac conduction disease. *SCN5A*, whose structure can be seen in figure 1, codes for the primary voltage-gated sodium channel isoform in human heart tissue will over 100 known variants accounting for cardiac arrhythmia phenotypes [10]. Cases of "sudden death" caused by an abnormal cardiac event in a seemingly healthy individual serves as frightening examples of the need for genomic analysis of proteins. New mutations of *SCN5A* are still being discovered, such as a recently discovered novel missense mutation (2352 G->A) in 2013 that had also been passed down to the patient's son [11]. Brugada syndrome is now known to be responsible for sudden unexplained death syndrome (SUDS). Given the crucial role that sodium voltage-gated channels play in cardiac physiology, the analysis of this protein family along with it's paralog, the voltage-gated calcium channel by MDS can aid in understanding how a gene like *SCN5A* evolved and in what sections mutation was allowed to happen in order to give rise to beneficial phenotypes, vs. the disease causing alleles. The voltage-gated potassium channels are the largest and most diverse family of ion channel found in virtually all organisms and are also in studying ion-channel evolution, given that the wide variety of potassium channels was caused by multiple duplication events [12].

METHODS

1. Software

MDS is used to analyze a collection of human voltage-gated ion channels as a proof-of-concept to show the powerful capabilities that MDS has. The bios2mds package is developed in

the Bioinformatics team at Integrated Neurovascular Biology Laboratory at the University of Angers. This package analyzes biological sequences by metric MDS with projection of supplementary data. It contains functions for reading multiple sequence alignment (MSA) files, calculating distance matrices from MSA files, performing MDS analysis and visualizing results. The MDS analysis and visualization tools can be applied to any kind of data. This package is used with *R*, a statistical programming language and software environment used for statistical computing, data analysis and graphics generation. The environment can be freely downloaded at http://www.r-project.org. Once *R* was opened, the *bios2mds* package was installed. Use of this program ideally requires previous experience with the *R* programming language but it is not essential. Basic operational steps done in *R* (installing *bios2mds*, opening the package, performing commands) are not laid out in detail but all the code used can be seen in the **R CODE** section, and a *bios2mds* manual written by the authors is available.

2. File Creation

Running the MDS analysis requires the creation of MSA files with a fasta (.fa or .fasta) extension. Since this filetype could not be created natively (via addition of a .fa suffix to a fasta formatted file) the packages sample MSA was simply overwritten with the custom MSA. Sequence data for human voltage-gated sodium, calcium and potassium (delayed rectifier subclass) was collected, and 1 multiple sequence was computed using the Clustal Omega program on the EMBL-EBI website. The output MSA was saved into a template .fa file by simply pasting the MSA (in fasta format) to the gpcr sample already present. The MSA created was named **all3human**, corresponding to: 1) **scnXa** (sodium), 2) **cacna1X** (calcium) and 3) **kcnaX** (potassium) and saved to the **msa** folder, as can be seen in figure 2. Once saved to the

appropriate folder (**Documents -> R -> win-library -> 3.1 -> bios2mds -> msa**). Custom csv color files corresponding to the color of each protein group were made in excel, and saved to the **csv** folder in the **bios2mds** folder by overwriting the provided gpcr file (**Documents -> R -> win-library -> 3.1 -> bios2mds -> csv**).

3. Performing the MDS Analysis

Visualizing MDS results is a 3 part process once the MSA files are obtained. First, a distance matrix of all the proteins must be made, which was calculated using a difference matrix. Second, the MDS computation was run, outputting the resulting eigenvalue decompositions. Third, a 2-dimensional plot of the MDS was created, followed by a 3-dimensional plot.

RESULTS

The resulting 2-dimensional and 3-dimensional plots of the MDS can be seen. The 2 dimensional plot appears to show a close spatial relationship between the 2 main groups of potassium channels, but upon moving to the 3-dimensional plot, a better relation can be seen and these 2 groups diverge off as their structural dissimilarity can be seen. These results are not surprising given that the voltage-gated potassium protein family is the largest and most diverse group of such proteins found in organisms. Groupings of such voltage-gated potassium channels align with the fact that these proteins have been found clustered in paralogous regions of mouse genomes [13]. MDS analysis reveals not only the structural, but evolutionary relationships of the voltage-gated ion channels. Given that these channels rely on a "paddle" motif of the critical

voltage sensing domain, and that these paddles can be transplanted across species and still retain their function as voltage-gated channels, this gives an insight into what regions of *SCN5A* give clinically manifested phenotypes of disease [14]. Future experiments should project voltage-gated orthologs of various species onto the human active space to see how the evolution of these channels has drifted from species to species.

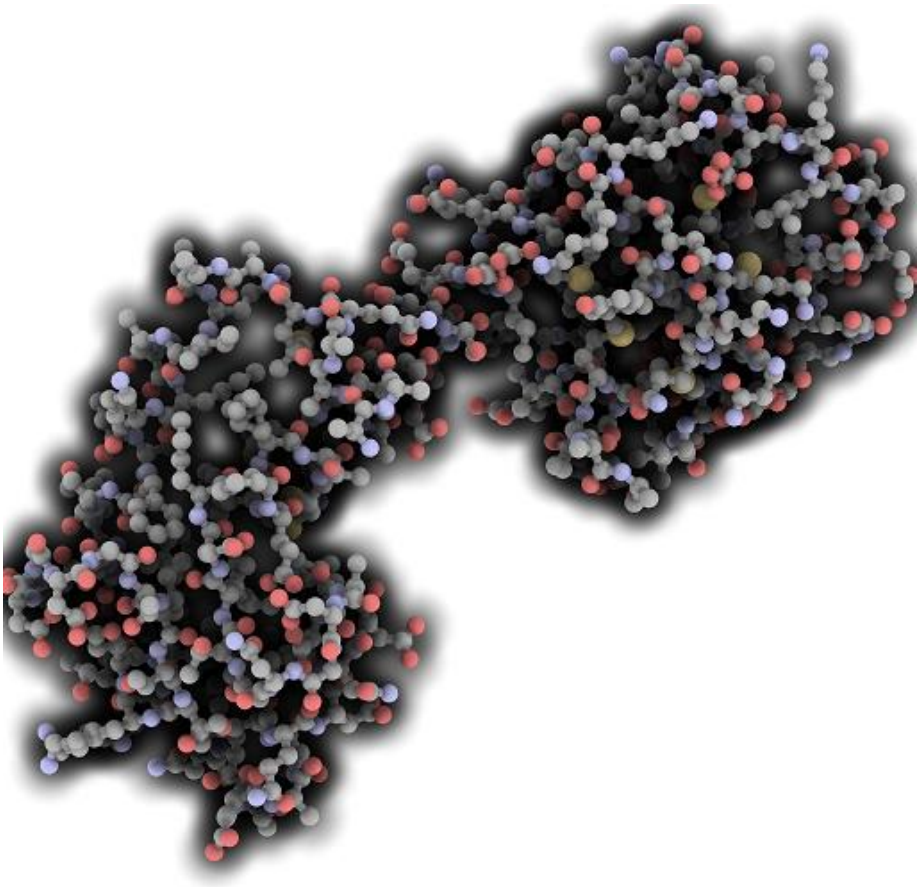

Figure 1: 3-D model of the SCN5A protein, rendered with QuteMol software using PDB entry 4DJC.

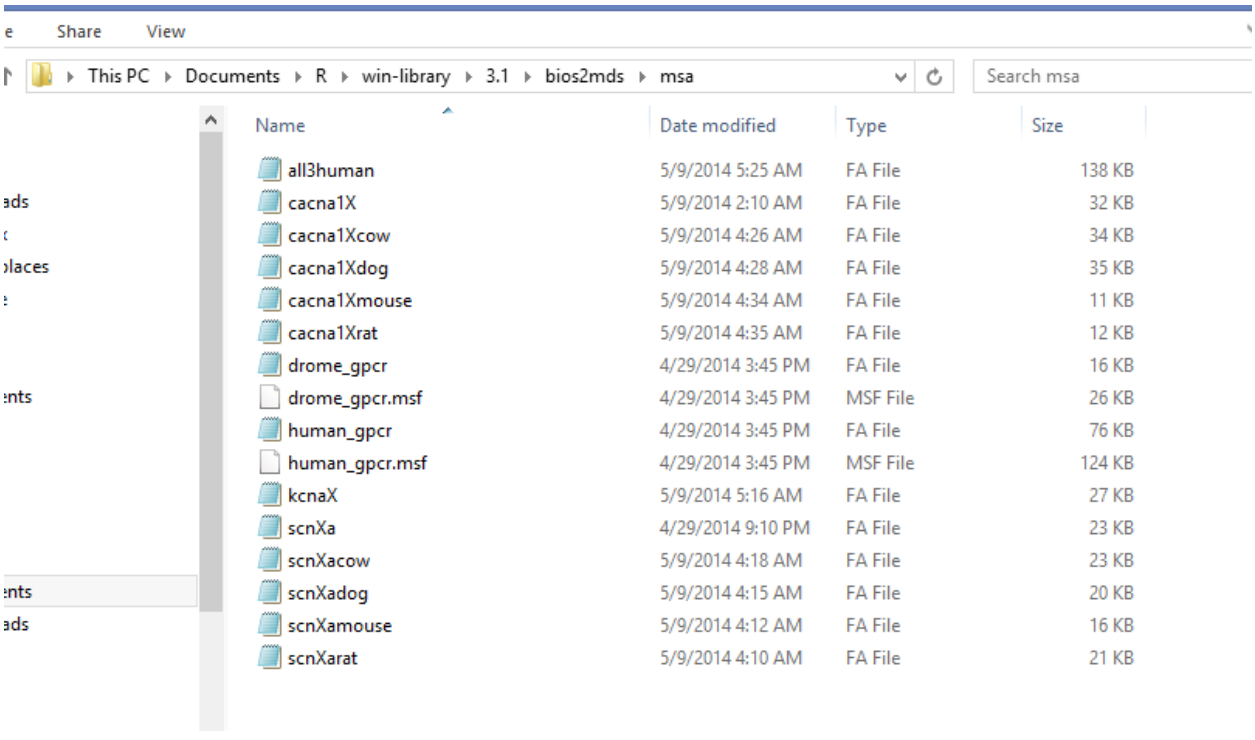

Figure 2: Directory where all the MSA files were saved.

R CODE & FIGURES

All code shown is for defining the active space. Code for the orthologs and supplementary is the same, just with file name changes so it is not shown to avoid redundancy.

```
> # Loading the bios2mds library to use with my sequences
> library(bios2mds)
Loading required package: amap
Loading required package: e1071
Loading required package: scales
Loading required package: cluster
Loading required package: rgl
```

```
> # Making the distance matrix of the human sodium, calcium and potassium alpha subunits
> aln <- import.fasta(system.file("msa/all3human.fa", package = "bios2mds"))
> mat.dis1 <- mat.dis(aln,aln)
> mat.dis1
        SCN1A SCN2A SCN3A SCN4A SCN5A SCN8A SCN9A SCN10A SCN11A SCN7A CACNA1S
SCN1A   0.000 0.078 0.099 0.205 0.250 0.153 0.151  0.302  0.358 0.341   0.611
SCN2A   0.078 0.000 0.079 0.197 0.248 0.155 0.145  0.299  0.353 0.334   0.599
SCN3A   0.099 0.079 0.000 0.205 0.243 0.167 0.157  0.299  0.354 0.342   0.603
SCN4A   0.205 0.197 0.205 0.000 0.244 0.212 0.209  0.290  0.358 0.367   0.609
SCN5A   0.250 0.248 0.243 0.244 0.000 0.263 0.263  0.255  0.339 0.391   0.619
SCN8A   0.153 0.155 0.167 0.212 0.263 0.000 0.192  0.307  0.354 0.353   0.614
SCN9A   0.151 0.145 0.157 0.209 0.263 0.192 0.000  0.306  0.361 0.319   0.604
SCN10A  0.302 0.299 0.299 0.290 0.255 0.307 0.306  0.000  0.349 0.418   0.626
SCN11A  0.358 0.353 0.354 0.358 0.339 0.354 0.361  0.349  0.000 0.463   0.617
SCN7A   0.341 0.334 0.342 0.367 0.391 0.353 0.319  0.418  0.463 0.000   0.652
CACNA1S 0.611 0.599 0.603 0.609 0.619 0.614 0.604  0.626  0.617 0.652   0.000
CACNA1C 0.611 0.609 0.612 0.622 0.621 0.623 0.609  0.638  0.636 0.661   0.238
CACNA1D 0.607 0.603 0.605 0.615 0.620 0.620 0.606  0.636  0.632 0.658   0.243
CACNA1F 0.621 0.613 0.615 0.613 0.624 0.626 0.615  0.637  0.626 0.669   0.276
CACNA1A 0.622 0.620 0.626 0.633 0.628 0.631 0.618  0.641  0.648 0.655   0.449
CACNA1B 0.628 0.627 0.631 0.639 0.633 0.632 0.627  0.643  0.652 0.660   0.453
CACNA1E 0.634 0.632 0.635 0.637 0.640 0.636 0.630  0.648  0.636 0.655   0.448
CACNA1G 0.641 0.640 0.638 0.638 0.661 0.651 0.645  0.649  0.641 0.664   0.619
CACNA1H 0.655 0.646 0.644 0.645 0.664 0.653 0.653  0.666  0.639 0.676   0.611
CACNA1I 0.646 0.637 0.636 0.646 0.651 0.647 0.644  0.648  0.642 0.667   0.612
KCNA1   0.790 0.797 0.804 0.803 0.802 0.792 0.796  0.814  0.813 0.824   0.784
KCNA2   0.786 0.791 0.792 0.797 0.794 0.789 0.790  0.802  0.803 0.817   0.788
KCNA3   0.819 0.824 0.829 0.816 0.815 0.819 0.813  0.837  0.836 0.826   0.789
KCNA5   0.805 0.809 0.813 0.809 0.825 0.807 0.807  0.824  0.826 0.830   0.794
KCNA6   0.799 0.803 0.809 0.813 0.812 0.809 0.797  0.820  0.827 0.821   0.782
KCNA7   0.770 0.778 0.784 0.778 0.770 0.780 0.771  0.780  0.798 0.803   0.799
KCNA10  0.806 0.812 0.818 0.818 0.833 0.822 0.813  0.839  0.846 0.835   0.780
KCNB1   0.778 0.776 0.783 0.788 0.801 0.787 0.778  0.788  0.787 0.806   0.821
KCNB2   0.784 0.778 0.786 0.808 0.810 0.792 0.780  0.791  0.800 0.812   0.815
KCNC1   0.788 0.792 0.796 0.798 0.801 0.796 0.785  0.791  0.814 0.834   0.801
KCNC2   0.829 0.829 0.831 0.842 0.844 0.840 0.831  0.836  0.860 0.869   0.836
KCNQ1   0.777 0.765 0.777 0.776 0.788 0.782 0.760  0.794  0.794 0.768   0.836
KCNQ2   0.784 0.768 0.778 0.791 0.798 0.791 0.765  0.809  0.807 0.781   0.807
KCNQ3   0.807 0.796 0.803 0.820 0.831 0.826 0.791  0.831  0.837 0.817   0.818
KCNQ4   0.798 0.794 0.793 0.810 0.822 0.821 0.780  0.828  0.826 0.821   0.805
KCNQ5   0.818 0.814 0.819 0.836 0.845 0.835 0.808  0.846  0.840 0.845   0.826
KCNH1   0.831 0.836 0.837 0.843 0.843 0.852 0.826  0.847  0.857 0.855   0.850
```

```
> # Perform an MDS on the distance matrix
> active <- mat.dis1
> mmds(active)
```

	SCN1A	SCN2A	SCN3A	SCN4A	SCN5A	SCN8A	SCN9A	SCN10A
SCN1A	0.000000	0.006084	0.009801	0.042025	0.062500	0.023409	0.022801	0.091204
SCN2A	0.006084	0.000000	0.006241	0.038809	0.061504	0.024025	0.021025	0.089401
SCN3A	0.009801	0.006241	0.000000	0.042025	0.059049	0.027889	0.024649	0.089401
SCN4A	0.042025	0.038809	0.042025	0.000000	0.059536	0.044944	0.043681	0.084100
SCN5A	0.062500	0.061504	0.059049	0.059536	0.000000	0.069169	0.069169	0.065025
SCN8A	0.023409	0.024025	0.027889	0.044944	0.069169	0.000000	0.036864	0.094249
SCN9A	0.022801	0.021025	0.024649	0.043681	0.069169	0.036864	0.000000	0.093636
SCN10A	0.091204	0.089401	0.089401	0.084100	0.065025	0.094249	0.093636	0.000000
SCN11A	0.128164	0.124609	0.125316	0.128164	0.114921	0.125316	0.130321	0.121801
SCN7A	0.116281	0.111556	0.116964	0.134689	0.152881	0.124609	0.101761	0.174724
CACNA1S	0.373321	0.358801	0.363609	0.370881	0.383161	0.376996	0.364816	0.391876
CACNA1C	0.373321	0.370881	0.374544	0.386884	0.385641	0.388129	0.370881	0.407044
CACNA1D	0.368449	0.363609	0.366025	0.378225	0.384400	0.384400	0.367236	0.404496
CACNA1F	0.385641	0.375769	0.378225	0.375769	0.389376	0.391876	0.378225	0.405769
CACNA1A	0.386884	0.384400	0.391876	0.400689	0.394384	0.398161	0.381924	0.410881
CACNA1B	0.394384	0.393129	0.398161	0.408321	0.400689	0.399424	0.393129	0.413449
CACNA1E	0.401956	0.399424	0.403225	0.405769	0.409600	0.404496	0.396900	0.419904
CACNA1G	0.410881	0.409600	0.407044	0.407044	0.436921	0.423801	0.416025	0.421201
CACNA1H	0.429025	0.417316	0.414736	0.416025	0.440896	0.426409	0.426409	0.443556
CACNA1I	0.417316	0.405769	0.404496	0.417316	0.423801	0.418609	0.414736	0.419904
KCNA1	0.624100	0.635209	0.646416	0.644809	0.643204	0.627264	0.633616	0.662596
KCNA2	0.617796	0.625681	0.627264	0.635209	0.630436	0.622521	0.624100	0.643204
KCNA3	0.670761	0.678976	0.687241	0.665856	0.664225	0.670761	0.660969	0.700569
KCNA5	0.648025	0.654481	0.660969	0.654481	0.680625	0.651249	0.651249	0.678976
KCNA6	0.638401	0.644809	0.654481	0.660969	0.659344	0.654481	0.635209	0.672400
KCNA7	0.592900	0.605284	0.614656	0.605284	0.592900	0.608400	0.594441	0.608400
KCNA10	0.649636	0.659344	0.669124	0.669124	0.693889	0.675684	0.660969	0.703921
KCNB1	0.605284	0.602176	0.613089	0.620944	0.641601	0.619369	0.605284	0.620944
KCNB2	0.614656	0.605284	0.617796	0.652864	0.656100	0.627264	0.608400	0.625681
KCNC1	0.620944	0.627264	0.633616	0.636804	0.641601	0.633616	0.616225	0.625681
KCNC2	0.687241	0.687241	0.690561	0.708964	0.712336	0.705600	0.690561	0.698896
KCNQ1	0.603729	0.585225	0.603729	0.602176	0.620944	0.611524	0.577600	0.630436
KCNQ2	0.614656	0.589824	0.605284	0.625681	0.636804	0.625681	0.585225	0.654481
KCNQ3	0.651249	0.633616	0.644809	0.672400	0.690561	0.682276	0.625681	0.690561
KCNQ4	0.636804	0.630436	0.628849	0.656100	0.675684	0.674041	0.608400	0.685584
KCNQ5	0.669124	0.662596	0.670761	0.698896	0.714025	0.697225	0.652864	0.715716
KCNH1	0.690561	0.698896	0.700569	0.710649	0.710649	0.725904	0.682276	0.717409

	SCN11A	SCN7A	CACNA1S	CACNA1C	CACNA1D	CACNA1F	CACNA1A	CACNA1B
SCN1A	0.128164	0.116281	0.373321	0.373321	0.368449	0.385641	0.386884	0.394384
SCN2A	0.124609	0.111556	0.358801	0.370881	0.363609	0.375769	0.384400	0.393129
SCN3A	0.125316	0.116964	0.363609	0.374544	0.366025	0.378225	0.391876	0.398161
SCN4A	0.128164	0.134689	0.370881	0.386884	0.378225	0.375769	0.400689	0.408321
SCN5A	0.114921	0.152881	0.383161	0.385641	0.384400	0.389376	0.394384	0.400689

```
> # Make a 2 dimensional MDS plot of the human sodium, calcium and potassium alpha subunits
> mmds.2D.plot(coloredmds, active.alpha=0.5, active.lab=TRUE, active.legend.pos="topright")
```

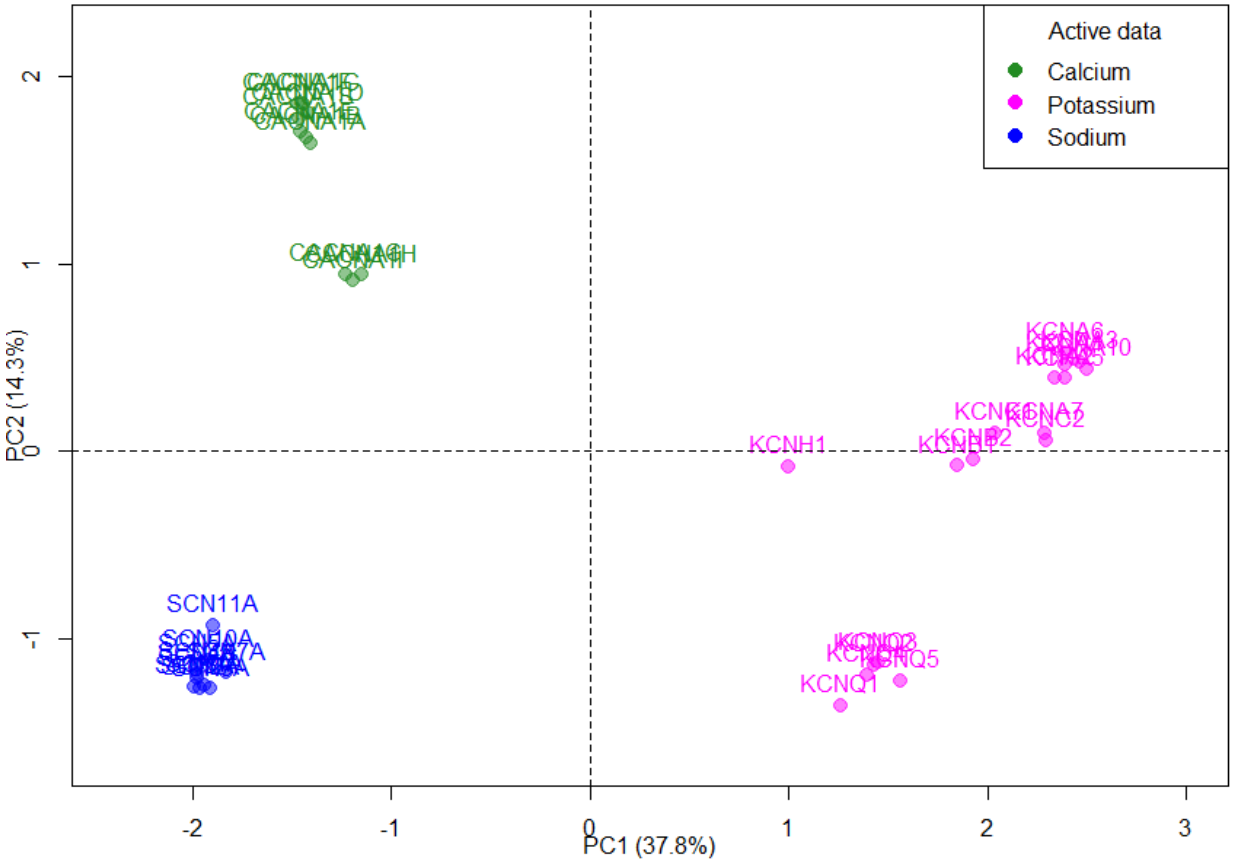

Figure 3: A 2-Dimensional MDS Plot of human calcium, potassium and sodium gated channels (all alpha subunit, potassium subfamily of delayed rectifiers) proteins. A simple difference matrix was used here to compute the distances since all proteins are human.

```
> # Make a 3D plot of the MDS just performed with spheres of a defined radius
> mmds.3D.plot(coloredmds, radius=0.05)
> bbox3d(shininess=0.5)
```

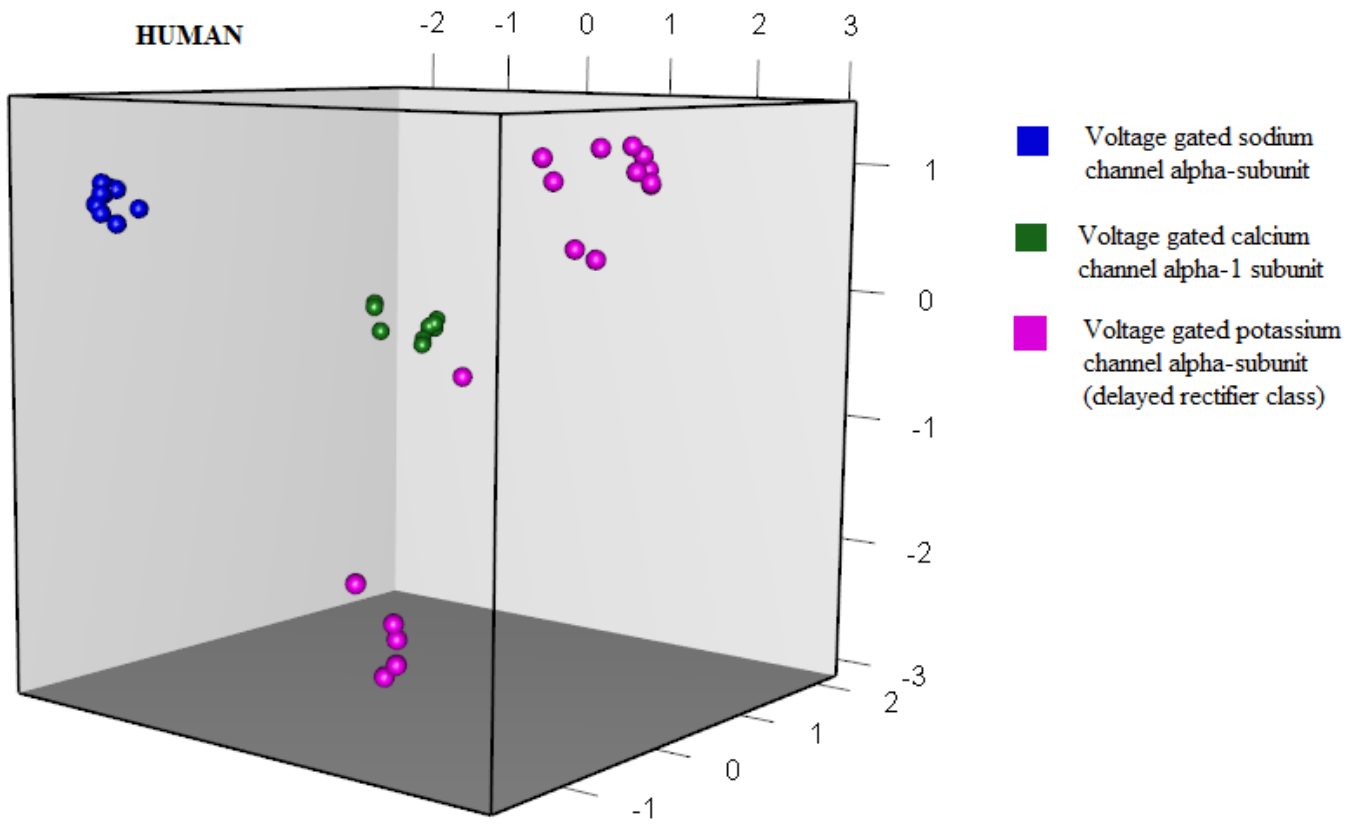

Figure 4: A 3-Dimensional Representation of the same data in figure 1. As more dimensions are added, the data becomes more representative of the real sequence space since fewer approximations need to be made. Better relations between these 3 protein families can be made. In the R program, this cube can actually be rotated to better see the spatial relationships between these protein families.

REFERENCES

1) Bornberg-Bauer E, Chan HS. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. Proceedings of the National Academy of Sciences. 1999;96(19):10689-10694.3) VMC PDF

4) Lesk, Arthur M. "5: Pattern Matching." *Introduction to Genomics*. Oxford: Oxford UP, 2012. 167-76. Print.

5) Pelé J, Bécu JM, Abdi H, Chabbert M. Bios2mds: an R package for comparing orthologous protein families by metric multidimensional scaling. BMC Bioinformatics. 2012;13(1):133.

6) Young. F. W. (1984). *Research Methods for Multimode Data Analvsis in the Behavioral Sciences.* H. G. Law, C. W. Snyder, J. Hattie, and R. P. MacDonald, eds.

7) Available at: http://www.codecogs.com/latex/eqneditor.php. Accessed May 15, 2014.

8) Higgins DG. Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. Bioinformatics. 1992;8(1):15-22.

9) Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics. 1971;27(4):857-.

10) Liu K, Hipkens S, Yang T, et al. Recombinase-mediated cassette exchange to rapidly and efficiently generate mice with human cardiac sodium channels. Genesis. 2006;44(11):556-64.

11) Sayeed MZ, Salam MA, Haque MZ, Islam AK. Brugada syndrome with a novel missense mutation in SCN5A gene: A case report from Bangladesh. Indian Heart J. 2014;66(1):104-7.

12) Jegla T, Marlow HQ, Chen B, Simmons DK, Jacobo SM, Martindale MQ. Expanded functional diversity of shaker K(+) channels in cnidarians is driven by gene expansion. PLoS ONE. 2012;7(12):e51366.

13) Lock LF, Gilbert DJ, Street VA, et al. Voltage-gated potassium channel genes are clustered in paralogous regions of the mouse genome. Genomics. 1994;20(3):354-62.

14) Alabi AA, Bahamonde MI, Jung HJ, Kim JI, Swartz KJ. Portability of paddle motif function and pharmacology in voltage sensors. Nature. 2007;450(7168):370-5.