Using Nate’s Terribly Complicated Scripts

**Machine Learning**

***Data:*** Manipulating the coordinate data created by Pengzhi Zhang

All scripts relating to this topic are stored in the directory /home/nate/Desktop/Scripts/matlab

* Read all comments in the scripts themselves for formatting and input instructions.
* all scripts leave the charge in the last column

Distancefromcoordinate.m: Takes in coordinates, spits out Euclidean distance from the calcium ion

AddDistancetoCoordinates.m: Takes in coordinate/Euclidean distance and spits out a 5-concatenation of atom number, x, y, z coordinates, and Euclidean distance

Clustering.m:Takes in any format with charge in the last column and splits the data into positively and negatively charged sets

waterorientation.m: Takes in point+distance from AddDistancetoCoordinates. Adds a binary feature based on the orientation of the water molecule: 1 if O is towards the calcium, 0 if away. Requires very specific formatting.

***Python Modeling:***

All scripts relating to this topic are stored in the directory /home/nate/Desktop/Scripts

* Read comments
* All scripts only need to have the input files replaced. Training data and prediction data can be loaded at will.

chickentester.py: named for the chicken calmodulin’s calcium charge originally predicted by the script, but works on any set of training and prediction data. Input training data into train and data to be predicted into predict

Deeplearn.py: training into train, prediction data into test. Uses H2O deep learning instead of random forest

ML.py: originally written by PZ, it outputs variable importance. Only takes training data into class\_df which is split randomly for validation.

MLnew.py: takes in only training data into train, randomly splitting the set and outputting the validation predictions as numbers as well as presenting a few statistics

PabloML.py: Written by Professor Pablo Rondon, this uses scikit learn random forest regressor from data loaded into filename

* None of these scripts seem superior to the other, as the performance is largely dependent on the input data set. They are all thrown together for different outputs and analysis.

Chickentester averages 100 models predictions, for example.

Contact Analysis

In the directory /home/nate/Desktop/Contactmaps

natesbackbone.txt: Replace output file destination. Load a structure with any number of frames in vmd and replace the n in ‘set mol n’ with, n, the molecule id of the trajectory of interest. It indicates backbone-backbone contacts only with hydrogens not included by default.

natessidechain.txt: same protocol as natesbackbone, but only indicates sidechain-sidechain contacts

natescontactanalysis.txt: same protocol as the others, but indicates all atom-atom contacts. Significantly slower.

Formatting Contact Analysis Outputs for Matlab script

Run the output file through the following bash command:

sort -V file.txt | uniq | cut -f2,3 -d: | sort -V | uniq -c | sed ‘s/^\s\*//’ | sed ‘s/:/ /g’ | awk ‘{print $2,$3,($1-1)/f >newfile.txt

where f is the number of frames in the trajectory

Notes:

* Generates the probability of contact in a given trajectory for each residue-residue pair
* 4Å cutoff
* 4 residue minimum sequence distance for contact consideration