Aim:

1. To create a comprehensive classifier for AD datasets (diagnoser, structure identifier, level predictor) , one which works on one dataset at a time, and another which trains on multiple, specially enriched datasets  
2. To use expression data from datasets to identify genes uniquely expressed in AD in neurofibulary tangles

Hypothesis

1. Linear classifiers and those meant to handle small datasets will perform exceptionally poorly on the larger datasets, though they may succeed on the smaller sizes. The Tree and related classifiers should perform better on the larger dataset, provided there's a significant difference in gene expression between AD and Non-AD patients in the structure.

2. If Neurofibulary Tangles do have a unique and significant expressive-role in AD, then there will be a significant number of genes uniquely expressed in the Tangles.

Method

|  |  |  |
| --- | --- | --- |
| Step | Procedure | Justification |
| I. | Enrichment | This cross-section of neurological data varies widely and needs to be handled in such a way that the data can be analyzed across datasets, as well as within the dataset itself (for classification). |
| A. | Classification Enrichment for Individual Analysis. We will call them A-Enriched datasets. These use the raw datasets. | Perform a first kind of enrichment to prepare for the classifier which will operate on one dataset at a time. This enrichment's results will be further enriched for more specific steps later on, and will include all datasets. |
| 1. | Remove Irrelevant Data for the classifier, such as ID numbers, data from other diseases, etc. | Some datasets contain data for other diseases, such as MS, Dementia, and other Neurodegenerative disorder. These samples need to be removed from datasets by hand and consultation with the GEOquery information page. |
| 2. | Perform some statistical significance analysis on the raw dataset, on the raw difference between expression levels in AD and non-AD. | This will help us determine thresholds for removing insignificant variables. |
| a. | Find the appropriate significance. |  |
| b. | (Covariance? , Variance?) |  |
|  |  |  |
| 3. | Determine a threshold for insignificant data by taking the highest Statistical Significance from the above values. | The higher value will help us avoid overfitting the models. If models are underfit, then take a lower measure as the threshold. |
| 4. | Using the above threshold, remove any insignificant genes. | Insignificant genes are those which fall into a certain degree of similarity. |
| B. | Classification Enrichment for Cross-Analysis. These use A-Enriched datasets unless otherwise specified, and will be called B-Enriched datasets. | Perform a second enrichment to prepare the datasets for classifiers which use multiple datasets |
| 1. | Isolate Incompatible Dataset Groups | Some A-Enriched datasets contain data and metadata which make them incompatible for immediate analysis. Noting these incompatibilities is important for the next step. |
| 2. | Compile compatible raw datasets into larger group-file. | These groups can be used as larger sample sizes for classifier testing, and will also be used for determining AD-significant genes, and genes relevant to tangles. |
| 3. | Remove Irrelevant data. | As above. |
| 4. | Calculate Threshold based on significance | As above. |
| 5. | Remove statistically insignificant data. The result is the B-Enriched dataset. | As above. |
| C. | AD Significant Gene Enrichment Type 1. These use B-enriched datasets. | We now perform a third type of enrichment, which will allow us to isolate genes generally significant is AD, specifically to Tangles. |
| 1. | Remove all Non-AD data. | This step may seem strange, but it's justified by our goal - to isolate highly expressed AD genes. We already removed the similarly expressed genes between AD and non-AD, so when we isolate AD genes, that's all that remains. |
| 2. | (optional) remove all data not relevant to mid-level AD. | The tangle dataset specifies its samples as "mid-level AD" patients. It may be necessary to remove incompatibility between datasets then. |
| 3. | Make sure Tangle is Enriched-A dataset. Compile all neurological structure datasets into one file, adding their samples as new samples. | This will allow us to compare the expression of significant AD-Tangle genes with significant AD genes in other neurological structures. |
| 4. | Prune genes which remain in the added datasets. | These genes are significant in other structures but not in Tangles or Non-Tangle structures. |
| 4. | Calculate variance of expression for each gene between all neurological structures | The variance of that gene shows us how much it is fluctuating between structures (we're only doing this with 'significant Tangle genes). |
| 5. | Remove statistically insignificant genes and place them into a separate dataset. | Hypothetically, this should leave us with genes which are highly dissimilar between neurological structures during AD, including Tangles. |
|  |  |  |
|  |  |  |
| II. | Classification | I aim to classify AD presence based on expression profile. |
| 1. | Prepare two sets of data: A-Enriched and B-Enriched datasets. These will be our objects of classification. | The A-Enriched datasets have isolated significant genes for the dataset by itself. The B-Enriched datasets isolated genes significant to compatible datasets (like entire structures). |
| 2. | Write classifiers which train and test on individual datasets, or install prefabricated classifier packages. Multiple classification methods will be attempted:  Definitely: Linear, Naive Bayes, Generalized Additive, Tree, Adaptive Regression, Neural Net, KNearest, Cluster, Forest Possibly: others... | I'm currently not too familiar with the mathematics which back up all of the models, so I've chosen a cross-section to at least become familiar with through this study. If I do end up writing one myself, I will probably only end up writing one, because I have a lot of other steps to take beyond writing and learning the internal machinery of all classifiers. |
| 3. | Choose a method for training and testing - perhaps do multiple methods:  Bootstrap Max Likelihood? | Each model may or may not favor a method of training and testing. It would provide more results to test multiple methods. |
| 4. | Apply each classifier to each dataset for 10 training and testing percentages:  10% test, 20% test, 30%test... etc | This will give me data to test against my hypotheses. This should be done by one R script, outputting individual result files for each dataset. |
| 5. | Wait for a long time. | Because it will take forever. |
| 6. | Compile and Analyze results. | I have to do this for my report. The results should show us the success of each classifier on the individual datasets, their runtime, etc, as well as the success of the classifiers on the compiled data. |
| III. | Analysis of Isolated Significant Genes in Neurofibulary tangles. |  |
| 1. | Perform analysis by isolating highly expressed genes which are highly expressed only for neurofibulary tangles. |  |
| 2. | This is done by taking the C-Enriched dataset and choosing the statistically significant genes which also belong to Tangles. |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |