**Mouse Model Projection Results**

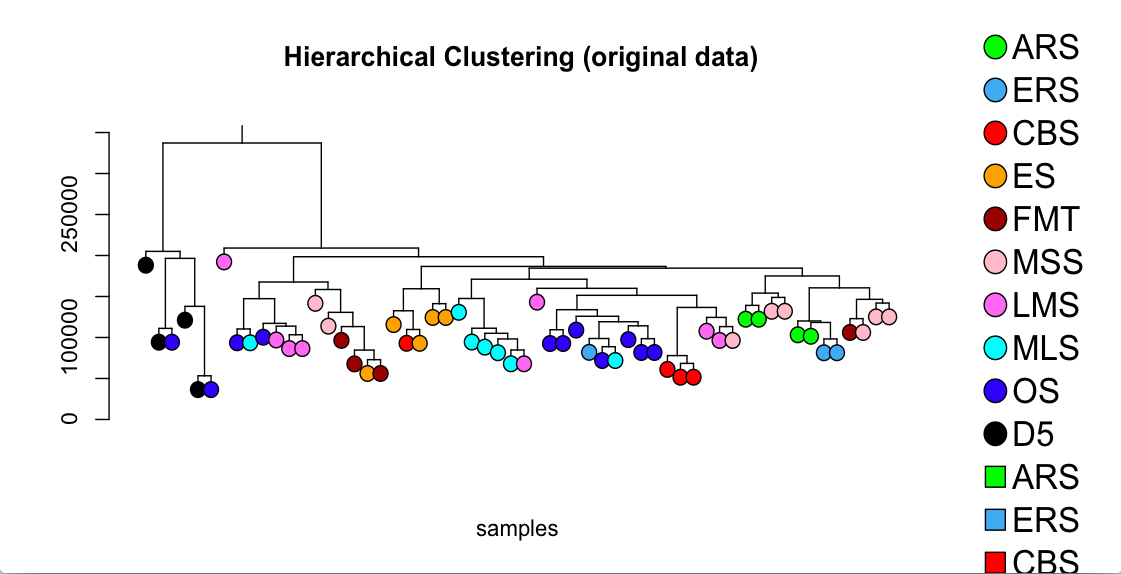
Tamayo et. al. had formulated a new algorithm using non-negative matrix factoring and pseudoinverses to project genomic data into metagenes, which could be used to classify subtypes of various subtypes, find clearer structure in genes using hierarchical clustering or even better gene set enrichment analysis. This project was focused on applying this algorithm to sarcoma data.

The sarcoma data comprised of 10 different classes with the following groupings:

* Alveolar rhabdomyosarcoma: 4 samples
* Embryonal rhabdomyosarcoma: 3 samples
* Chondroblastoma: 4 samples
* Ewing’s sarcoma: 5 samples
* Fibromatosis: 4 samples
* Monophasic synovial sarcoma: 8 samples
* Leiomyosarcoma: 7 samples
* Myxoid liposarcoma: 7 samples
* Osteosarcoma: 11 samples
* Unknown tumor D5: 6 samples

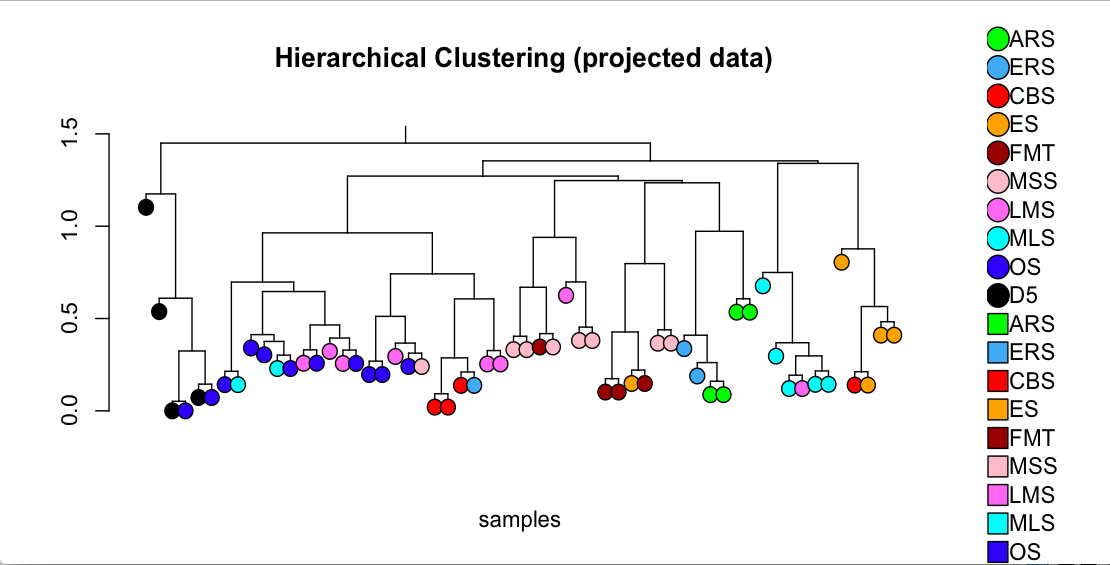
Using bioMart packages in R and Ensembl annotation datasets, all samples were first converted to gene expression levels. Any genes that were not found in any other samples were eliminated. Finally, datasets were merged and split into training and testing datasets through equal splitting (if the total number of samples were odd, the majority of the samples would go to the training dataset).

After running the algorithm, 3 plots were created: hierarchical clustering of the original data, hierarchical clustering of the projected data, and a heatmap of the metaganes. The following is a brief analysis of these plots:



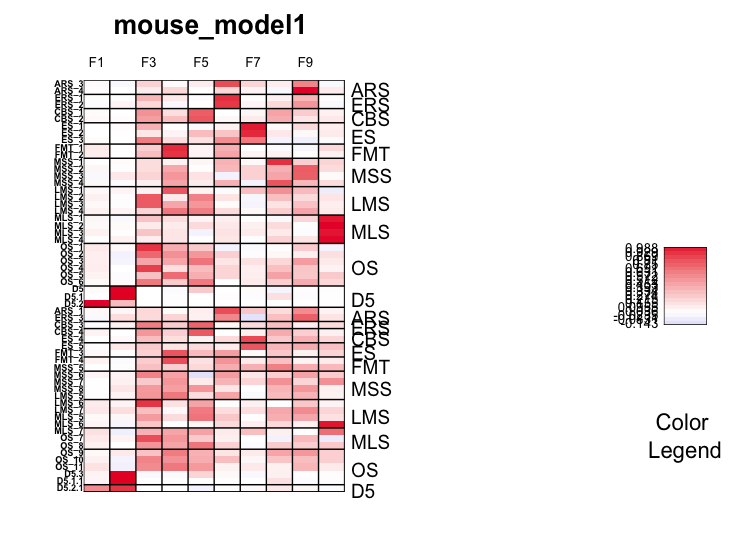
In the hierarchical clustering of the original data, we notice the following:

* Clusters are not homogenous with a single color. Overall, they are loosely clustered
* Certain subtypes, like chondroblastoma and myxoid liposarcoma are well clustered
* The unknown tumor is separated from the other subtypes, but is most closely clustered with osteosarcoma.



In the hierarchical clustering of the projected data, we notice the following:

* Clustering is somewhat more homogenous (refer to Ewing’s sarcoma and osteosarcoma)
* Previously defined clusters, such as chondroblastoma, have broken up
* The unknown tumor is still clustered away from other subtypes, but is most closely clustered to osteosarcoma



In the heatmap of the metagenes (numbered F1 to F10 on the top), the major trend that appears is that the metagene profile of D5 is totally different from other subtypes (as seen by the heatmap of F1 and F2).

Furthermore, the metagenes were used to predict on the test dataset. The following is a table of the performance of the algorithm

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Class | Count | No Call | No Call (%) | Real Error | Real Error (%) | Correct Call | Correct Call |
| ARS | 2 | 1 | 0.5 | 1 | 0.5 | 0 | 0 |
| CBS | 2 | 1 | 0.5 | 0 | 0 | 1 | 0.5 |
| D5 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| ERS | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| ES | 2 | 1 | 0.5 | 0 | 0 | 1 | 0.5 |
| FMT | 2 | 2 | 1 | 0 | 0 | 0 | 0 |
| LMS | 4 | 2 | 0.5 | 1 | 0.25 | 1 | 0.25 |
| MLS | 3 | 1 | 0.33 | 2 | 0.66 | 0 | 0 |
| MSS | 4 | 3 | 0.75 | 1 | 0.25 | 0 | 0 |
| OS | 5 | 1 | 0.2 | 4 | 0.8 | 0 | 0 |
| Total | 26 | 13 | 0.5 | 10 | 0.39 | 3 | 0.12 |

The table above suggests that the algorithm did not perform well in classification tasks. Furthermore, a high percentage of no calls indicates that the algorithm cannot decide classes for half of the samples.

The performance of the algorithm could be attributed to the following:

1. Low amount of samples for training may have resulted in weaker metagene projections
2. Originally, all subtypes, except for the unknown tumor, were encoded in transcript format. Although it is possible to convert the samples from transcripts to genes, the expression level may change as a result, which was not accounted for
3. Many genes were lost as the same genes had to be present in all samples. This information could have been used to diagnose certain samples in a better fashion, but could not be used for other subtypes
4. No post-projection normalization was used as it often eliminated entire classes due to the low amount of training samples
5. A stochastic process of finding the optimal number of metagenes was not used
6. The no-call threshold may have been too high. Lowering this threshold may result in more correct calls.