|  |
| --- |
|  |
| \section{Clustering with a single similarity measure and sequences} |
|  |  |
|  | To reiterate, the SCOPe database uses a mixture of both manual and automatic methods to group the present domains into its hierarchy. Not only does this classification use protein structures, it also uses their sequences to obtain the hierarchies' labels. Given this, in an attempt to obtain better results in terms of external evaluation metrics, we combined the sequence alignment distance matrix with the structure alignment ones. In the table below, we can see the results of this process. |
|  |  |
|  | The obtained results by clustering with sequence were somewhat unexpected, as we hoped it would improve performance by filling some ‘information gaps’ that might have been missing. The similarity measures' distance matrices were combined one by one with the sequence distance matrix by taking complementary percentages of both of them. After this process was finished, we went through the results to find which combination produced the highest cluster evaluation metrics. However, after analyzing the results we  saw that by taking even a small percentage of the sequence matrix into account only decreases performance, contrary to what was expected. Once again, this is most likely due to the fact that at the superfamily level, proteins do not share many similarities sequence wise and it becomes difficult to detect traces of homology through it. |
|  |  |

|  |
| --- |
|  |
| \section{Clustering with two similarity measures} |
|  |  |
|  | Given that the use of sequences does not translate into an increase in performance, from this point on the experiences made were only based on the structure aspect of the proteins. |
|  |  |
|  | This next step of the experiment focused on one of the main goals of this work: to explore the combinations of similarity measures. In this phase of the experiments, it was expected that we could improve the internal clustering metrics, since we aim to counter the disadvantages of other similarity measures with other ones or at least complement the each other’s data.  At this stage, we were hoping that most of the combinations that were experimented with would be improved in relation to the single measure experiment. In the figure below, the results of this experiment is displayed.  The results that were obtained with the hierarchical algorithm were as expected. This experiment does confirm that there are in fact small improvements in the clustering evaluation metrics when combining matrices. However, such improvements are mostly exclusive to the RMSD classifications as they were the most affected when used in combination with other measures. In most cases, it was shown that RMSD is massively outweighed by other similarity measures, since the classifications that are improved only take into account around 5% of the values present in its distance matrix. |
|  | One interesting observation regarding the RMSD classifications is that they seem to be irrelevant when combined with either Maxsub or TM-score. This is evidenced by the fact that when RMSD is combined with either one of these two, the respective cluster evaluation metrics tend to be nearly equal to their metrics when clustered by themselves. On the other hand  Furthermore, the other measure combinations also show improvements when compared to the single matrix clustering metrics. |
|  |  |
|  | The results that were obtained with the hierarchical algorithm were very favorable. Even though, the weights for the different pairs of similarity measures vary across samples, we can see that we are able to complement RMSD with other measures in order to improve both internal and external clustering metrics. Besides this, |
|  |  |
|  |  |
|  |  |
|  | \section{Clustering multiple similarity matrices} |
|  |  |
|  | The final experiment was made using the several similarity distance matrices. In order to determine what percentage to take from each of them, a genetic algorithm was used, which saved us a lot of time since it would not be necessary to cluster the matrices with every single combination of weights. |
|  |  |