1. SIMILARITY METRIC

Jaccard similarity, also known as the Intersection over Union, is a measure of similarity and diversity across two sets.[1] It is calculated as the number of unique items in common between the two sets, divided by the total number of unique items:

$$J(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Python implementation:

$$\frac{set.intersection(*\,[set(residues_a), set(residues_b)])}{set.union(*\,[set(residues_a), set(residues_b)])}$$

I chose this metric because of the logical expectation that active sites composed of similar lists of amino acid residues will have similar biological properties. However, this method does not take into account the number of times a particular residue appears in the active sites being compared (which could feasibly have a large impact on the kinds of proteins that bind to the site); additionally, it penalizes all residue differences equally, without accounting for the relative similarities of those residues (e.g. two active sites with different membership states of polar N and polar H might be expected to have more similar properties than two active sites that differ in membership of polar N and nonpolar L).

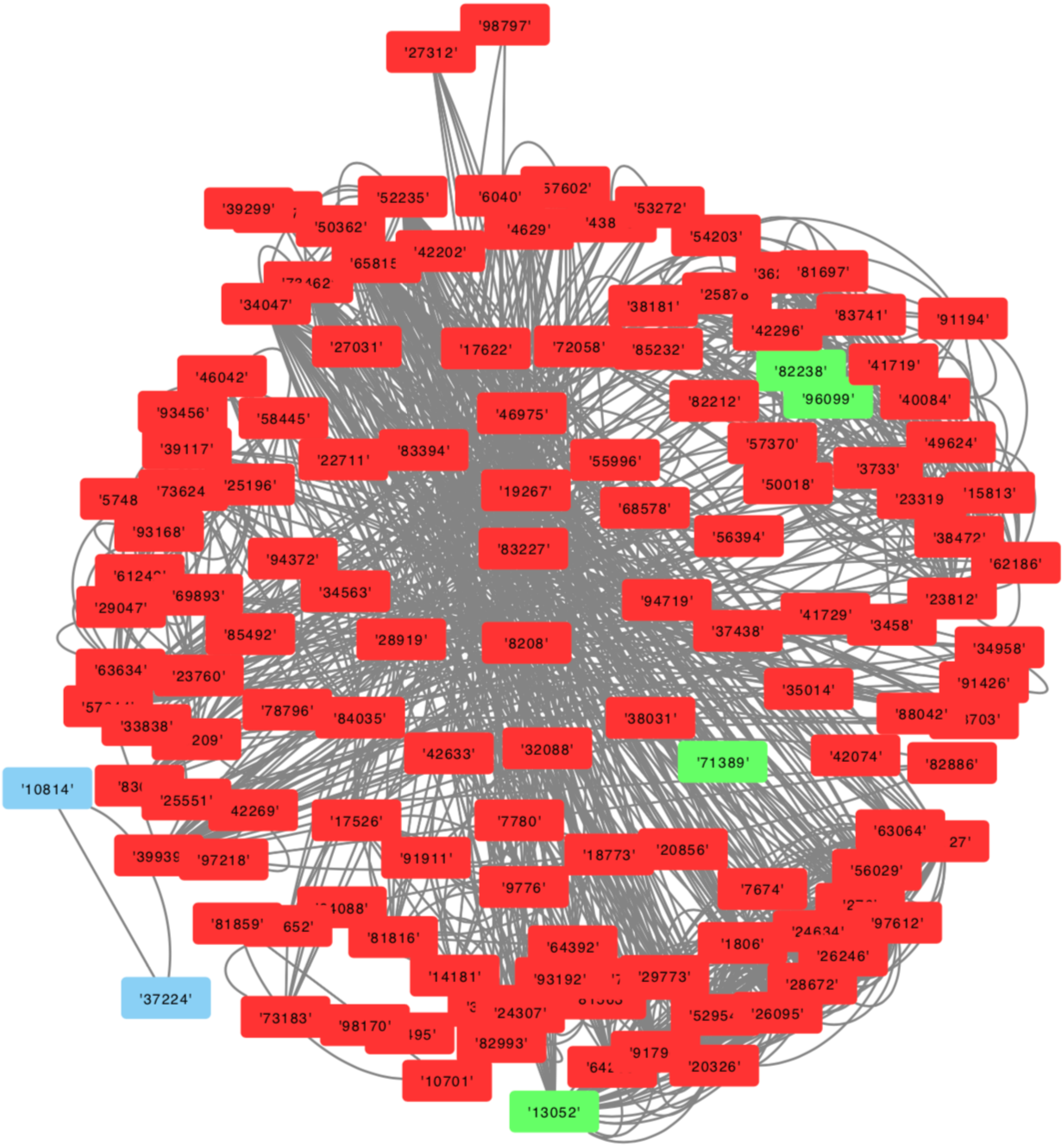
2. COVERING PARTITIONING ALGORITHM

The partitioning algorithm implemented here is a version of a covering algorithm, which (unlike most partitioning algorithms) does not require a predetermined number of desired clusters.[2] Instead it relies on a pairwise similarity distribution and threshold percentile cutoff to determine cluster separation. Since I had no prior knowledge of the biological significance of the given active sites, I chose this method because it did not require input of *k* desired clusters.

However, the covering algorithm does require a percentile cutoff to guide the threshold similarity value. To determine the ideal cutoff, I calculated the average silhouette score (discussed in more detail in Section 4) for ten rounds of clustering at percentile cutoffs ranging from 0 to 20 in steps of 0.5. The 15.0th percentile was selected as it led to clusterings with the highest silhouette score.

As another caveat, covering algorithms are subject to change based on the initial condition (i.e. the order in which the active sites are observed), so the input list was shuffled before running. In general, it is desirable to run the function several times and create a consensus clustering to “average” the results together; in the interest of time, I have only informally confirmed that the clusterings were not radically different from one another by computing the Rand Index (discussed further in Section 5) between results (mean R = 0.92, SD = 0.04, n = 10).

A visualization of the resulting partition is shown below. Edges are weighted by the Jaccard similarity score (only edges of J > 0.8 are shown), and nodes represent PDB active sites colored according to cluster membership. This particular similarity metric/partitioning algorithm combination tends to create a single large cluster that envelopes most elements.

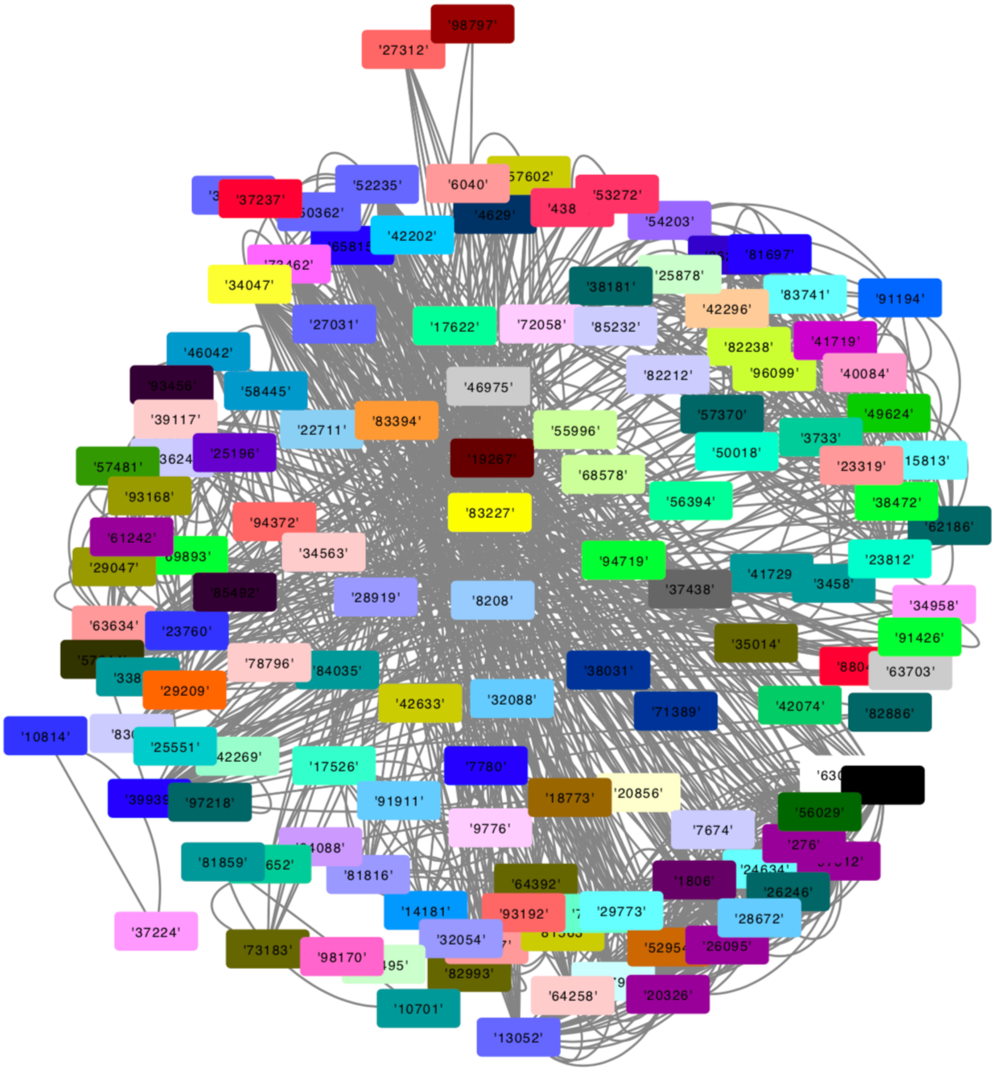


3. AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

In agglomerative hierarchical clustering, each active site begins as its own individual cluster. The two clusters with the highest similarity are merged into a single cluster, and similarity scores are calculated for the new cluster. This is performed iteratively until there is only a single cluster.

To answer the question of where to partition the resulting dendrogram, I have calculated the average silhouette score at each level of the agglomeration. The clustering with the highest score is accepted as the clustering result.

However, what if multiple possible clusterings have the same average silhouette score? I have examined both the first-encountered maximally-scored clustering (i.e. the clustering that performed maximally with the greatest number of clusters) and the last-encountered maximally-scored clustering (i.e. fewer number of bigger clusters). Based on this particular dataset and metrics, the ideal hierarchical partitioning creates 111 distinct clusters, with no ties for any alternative partitions (see below).



4. SILHOUETTE SCORE FOR EVALUATING CLUSTERINGS

The silhouette score is a measure of how well an element fits in with its cluster (cohesion) and how different it is from all other clusters (separation), given by:

$$\frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where b(i) indicates the similarity to items in the same cluster and a(i) indicates similarity to items in different clusters.[4] For a single cluster, the silhouette scores of each element can be averaged to produce a cluster silhouette, and a silhouette for the entire clustering is calculated as an average of all these silhouette scores. The silhouette score is on a scale of -1 (your clusters are deliberately scrambled) to 1 (each cluster is perfectly cohesive to itself and separate from all others).

The different clustering algorithms were performed and the silhouette coefficients calculated for N = 10.

Average Rand Index Hier: 0.9969
Standard Deviation Hier: 0.0003
Average Rand Index Part: 0.990
Standard Deviation Part: 0.005

Neither clustering performs significantly better, although the results are extremely different in terms of cluster size. This is likely due to the fact that parameters in both of the algorithms (threshold cutoff for the partitioning algorithm and cut height for the hierarchical cluster) were optimized for the silhouette score.

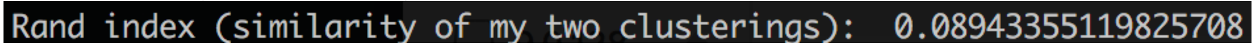
5. RAND INDEX FOR COMPARING CLUSTERINGS

The Rand Index is an external method for measuring the similarity of two clusterings. It is calculated as shown below:[5]

$$\frac{a + b}{a + b + c + d}$$

where a = number of pairs of elements that were grouped in the same cluster in both clusterings, b = number of pairs of elements grouped in different clusters in both clusterings, c = number of pairs of elements that were grouped in the same cluster in Clustering A but separately in Clustering B, and d = complement of c.

The result of the comparison between my covering partition and hierarchical cluster found that they are extremely different:



```
Rand index (similarity of my two clusterings):  0.08943355119825708
```

6. BIOLOGICAL MEANING

Examining the out-clusters in the partitioning result and the biggest clusters in the hierarchical dendrogram result showed no clear biological significance (based on charge, volume, or surface area). I believe that the similarity metric must be improved before meaningful results can be found. Additionally, the method for optimizing cluster quality should not be the same metric used to assess the success of the clustering, as I have done here with the Silhouette Score.

REFERENCES

1. Jaccard, Paul (1912). "The distribution of the flora in the alpine zone", New Phytologist, **11**: 37–50, doi:10.1111/j.1469-8137.1912.tb05611.x.
2. Pegg, Scott. Clustering lecture at UCSF, 01/24/2018.
3. Prosperi et al. (2011). Threshold Bootstrap Clustering.
4. https://cs.fit.edu/~pkc/classes/ml-internet/silhouette.pdf
5. W. M. Rand (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association*. American Statistical Association. **66** (336): 846–850. doi:10.2307/2284239