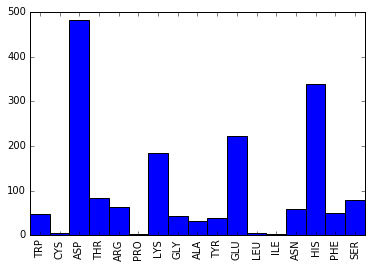
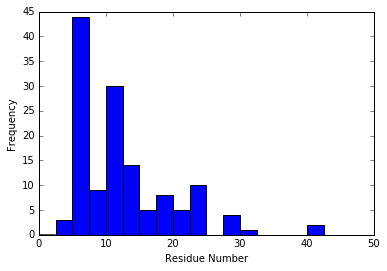
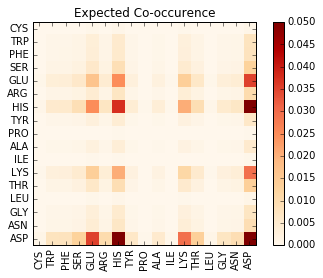
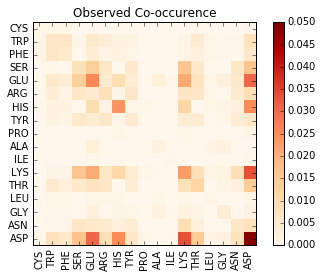
**Similarity Metric Description and Justification**

Before considering any similarity metrics, I summarized the data to look at global properties:



*Figure 1: Summary of active site dataset properties – the distribution of the total number of amino acid residues present in the active site (left) and the frequency of different residue types across the dataset (right).*

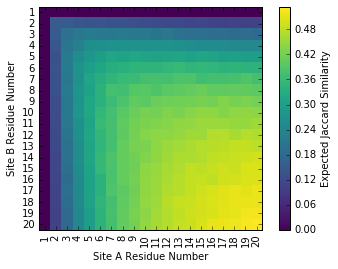
Based on observed frequencies of different residue types, it is possible to predict the probability of co-occurrence under the assumption of independence. If that prediction were similar to observed frequencies of co-occurrence, then amino acid composition would offer little information about the active site. However, there are dramatic differences between expected and observed co-occurrence of different amino acids.



*Figure 2: Observed and expected (under assumption of independence) co-occurrence of residue types in active sites in the dataset.*

To analyze the active sites, I chose to start with a simple similarity metric based on amino acid composition: the Jaccard similarity of the amino acid types present in the active site. Jaccard similarity is the ratio of residue types present in both active sites to the total number of residue types present in both active sites. For example, if active site A contains lysine, histidine, and aspartate and site B contains lysine, aspartate, and serine, their Jaccard similarity is 0.5.

One potential disadvantage of this metric is that the expected degree of similarity for randomly generated active sites varies with their size. Larger active sites are more likely to be similar to one another by pure chance. To understand this bias, I randomly generated active sites of different sizes using the observed distribution of residue types in the data and calculated the similarity of these randomly generated active sites.



In order to better distinguish larger active sites from one another, they should have a higher Jaccard similarity to be considered more similar than expected. So in order to account for these expected baseline similarities, I use for my similarity metric a ‘corrected’ Jaccard similarity:

where is the expected Jaccard similarity under independence. Given the complete set of residues from all active sites R = {r1,r1,r1,r2,r2,r2,r3,r4,r5,…}, the expected value, , of two randomly generated active sites of sizes and can be found by:

**Partitioning Algorithm**

I implemented a k-medioids algorithm to cluster the active site data. It begins by randomly choosing k data points as centers, then assigns all other data points to the ‘nearest’ center (where nearest means the center with the largest similarity). Within each cluster, the data point that has the maximum average similarity to all other data points within the cluster becomes the new center. All other points are then assigned to their new nearest center. This process is repeated until the clusters converge to a single answer.

**Hierarchical Algorithm**

**Clustering Quality Metric**

**Comparison of Clustering Algorithms**

**Biological Significance of Results**