# Cluster Utility: A New Metric for Clustering Biological Sequences

Jason Lee[1] and Sun Kim[1,2]
[1]School of Informatics, [2] Center for Genomics and Bioinformatics
Indiana University - Bloomington, IN 47404, USA

## 1. Introduction

Sequence clustering problem is different from traditional clustering problems in that the features of sequences are not observable and sequences cannot be placed in a metric space, which most clustering algorithms assume. The most widely used approach is to build a sequence graph using the all-pairwise sequence comparison data and to use the graph to generate clusters of sequences. Like other clustering problems, a metric to evaluate results from a sequence clustering algorithm is needed, but the metrics for traditional clustering problems are not readily applicable due to their metric space assumption. We propose *Cluster Utility* (CU), a metric that is based on consideration of similarity within a cluster and difference between clusters without metric space assumption. CU showed a very high correlation with the quality index. CU scales very well with data size and its strong correlation with quality index was nearly invariable regardless of data size change. CU can be used in two ways: to guide sequence clustering algorithms and to evaluate clustering results.

## 2. Cluster Utility

Given a candidate set of clusters $\{C_1, C_2, ..., C_k\}$ from a parent cluster $C$ and its associated graph $G$, graphs $G_i$ are generated for each candidate $C_i$. Let $I_{C_i}(s)$ denote a set of vertices that belong to a cluster $C_i$, among those adjacent to sequence $s$. Let $O_{C_i}(s)$ denote a set of vertices that do not belong to a cluster $C_i$, among those adjacent to sequence $s$. Both $I_{C_i}(s)$ and $O_{C_i}(s)$ are computed using the parent graph $G$ and not $G_i$. For a set of vertices or sequences $S$, $|S|$ denotes the size of the set. $R_{C_i}(s)$ represents the reachability of a sequence $s$ within a cluster $C_i$ and is defined as $|I_{C_i}(s)|/|C_i|$.

Let $L_{C_i}(s)$ denote $(|I_{C_i}(s)|+1)/(|O_{C_i}(s)|+1)$, which is the ratio of the number of vertices adjacent to $s$ within the cluster $C_i$ to the number of vertices adjacent to $s$ outside the cluster $C_i$,

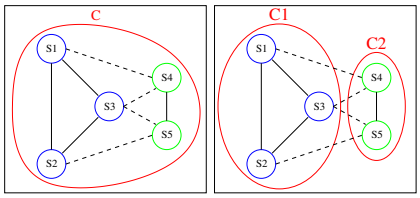
Note that we have added a pseudocount of one to both the numerator and the denominator to prevent division by 0, when $O_{C_i}(s)$ is empty.

Cluster utility $CU(C_i)$ of a cluster $C_i$ is defined as

$$CU(C_i) = \frac{\sum_{s \in C_i}(log(L_{C_i}(s)) \times R_{C_i}(s))}{|C_i|}$$

For a set of clusters $\{C_1, C_2, ..., C_k\}$, the cluster utility is the sum of cluster utility of each $C_i$, i.e, $CU(\{C_1, C_2, ..., C_k\}) = \sum_{i=1}^{i=k} CU(C_i)$.

**Example:** Let us compute the CUs for $C_1$ and $C_2$.




For $s_1$, there are two edges within $C_1$ and one edge outside $C_1$, $I_{C_1}(s_1) = 2$ and $O_{C_1}(s_1) = 1$, thus $L_{C_1}(s_1) = (2+1)/(1+1) = 1.5$. The reachability of $s_1$ is $R_{C_1}(s_1) = 2/3 = 0.667$. The numbers for $s_2$ and $s_3$ are computed similarly. Then $CU(C_1) = \frac{(0.27)*0.67+(0.27)*0.67+(0.00)*0.67}{3} = 0.180$ and $CU(C_2) = \frac{(-0.20)*0.50+(-0.20)*0.50}{2} = -0.203$. The final CU value for $C$ is $CU(C) = 0.180 + (-0.203) = -0.023$ which indicates that splitting is not desirable.

**Guiding clustering algorithms:** Let us consider the following example to illustrate how CU can be used to guide a sequence clustering algorithm. The data set includes four classes COG0002, COG0160, COG1959, COG3000 with 29, 79, 42, 17 sequences respectively. We have a choice between three different clustering results at three different cutoff values for pairwise similarity scores, 90, 100, and 110.

```
CU= -0.0829277 at cutoff 90
        c0 with sz= 17 : cu = 1.1473
        c1 with sz= 2 : cu = -2.37787
```

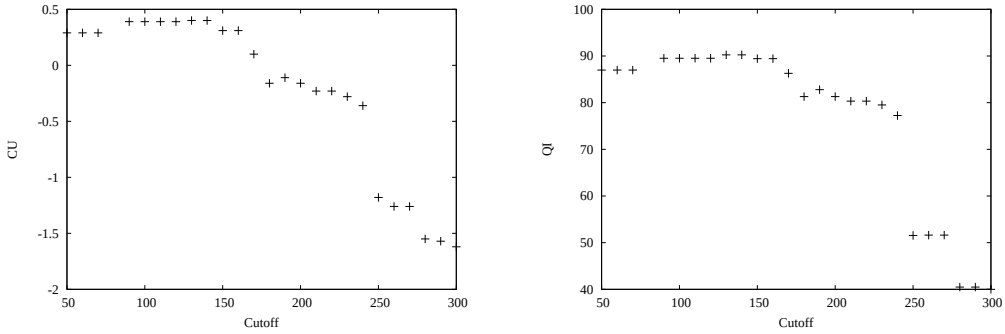



**Figure 1. CU (left plot) and QI (right plot) behave similarly for clustering results at different cutoffs.**

```
        c2 with sz= 150 : cu = 1.14764
CU= 6.54865 at cutoff 100
        c0 with sz= 29 : cu = 2.48778
        c1 with sz= 79 : cu = 3.32965
        c2 with sz= 42 : cu = 0.163948
        c3 with sz= 17 : cu = 0.567279
CU= 2.92823 at cutoff 110
        c0 with sz= 29 : cu = 2.32241
        c1 with sz= 79 : cu = 3.15243
        c2 with sz= 4 : cu = -0.794126
        c3 with sz= 2 : cu = -1.95099
        c4 with sz= 38 : cu = -0.0155198
        c5 with sz= 17 : cu = 0.214022
```

Which one do we choose? This decision is guided by comparing CUs and simply selecting the one with the highest CU value. A higher positive CU value implies a better clustering result since it means that more intra-cluster connectivity than inter-cluster connectivity. Thus we can choose the one at the cutoff of 100 easily.

We have implemented CU in our clustering algorithm BAG [1]. To distinguish it from the old version, we call the new version CU-BAG and the old version BAG. The performance of CU-BAG was tested with the COG database (2001 version) containing 77,114 sequences from 43 completely sequenced genomes. There are 3,311 different families and their family sizes vary widely from 2 to 806 sequences, which is a challenging clustering problem. The number of sequences in incorrect clusters reduced from 5687 sequences (BAG) to 1509 sequences (CU-BAG).

**Evaluating different clustering results:** Different sequence clustering algorithm can produce different clustering results. In addition, a clustering algorithm can produce different clustering results at different cutoff values for pairwise similarity. We can use CU to choose which one from different clustering results by simply selecting one with the highest CU value.

To test how good CU reflects the goodness of a clustering result, a data set comprised of three COG families, 0001, 0331, and 0523, was clustered by using BAG with cutoff values from 50 to 300 with an increase of 10, and CU and QI (quality index) values were measured. QI is an index to measure how many sequences are correctly clustered without fragmentation (all correct clusters except the maximal one are considered as fragmented). Figure 1 (left plot) depicts how CU changes according to the cutoff value change. Figure 1 (right plot) shows the relationship between the cutoff value and QI. QI varies as cutoff values changes. QI tends to be stable with respect to the cutoff value change over a range of approximately 50. The maximum QI is attained at the cutoff value of 140. The shapes of the two plots were surprisingly similar. Relative distribution of QI and CU points were the same, and maximum CU corresponded to maximum QI,

## 3. Discussion

It has been shown that CU can be used either to guide sequence clustering algorithms or to evaluate clustering results. CU exhibits several important properties: correlation with clustering quality, scalability, and insensitiveness to complexity of the input data. Comparative study demonstrated that CU far exceeds other competing indices in correlation with the clustering quality. Future work includes theoretical study of the behavior of CU as well as the improvement of CU.

## References

[1] S. Kim. Graph theoretic sequence clustering algorithms and their applications to genome comparison. In J. T. L. Wang, C. H. Wu, and P. Wang, editors, *Computational Biology and Genome Informatics*, pages 81–116. World Scientific, 2003.