A Framework Towards Efficient and Effective Protein Clustering

Wei Wang [1], **Jiong Yang** [2]

Keywords: protein sequence analysis, sequence clustering.

During recent years, many efforts have been carried out on analyzing protein sequences. Research in this area focuses on modeling and classifying protein families based on function, structure, and homology. Despite the differences in model definition and/or algorithm details, a major assumption of previous work is that the protein families are known in advance and the protein sequences used to build the classification model are properly labeled with the corresponding families. In this paper, we are studying a slightly different but more challenging problem — clustering unlabeled protein sequences into groups/families automatically.

Clustering has been widely recognized as a powerful technique in statistics and computer science, and has been studied extensively during recent years. The major goal of clustering is to create a partition of objects such that objects in each group have similar features. The result can potentially reveal unknown object groups/categories that may lead to a better understanding of the nature. Most research on clustering focused on numerical domains and assumed metric space in many cases. In many of these models, the similarity between objects can be defined in a clear and relatively straightforward way. Unfortunately, in the context that we address here, each object is a protein sequence and there is not obvious similarity measure that can be applied directly. This is in part because (1) each protein sequence is composed of symbols (i.e., amino acids) of non-numerical nature; (2) the length of a protein sequence can easily run up to thousands; and (3) different protein sequences may be of vastly different lengths. It is not feasible to map a protein sequence into a point in a high dimensional metric space and then utilize existing clustering algorithms developed on metric space. One possible approach is to use the edit distance between each pair of sequences as the distance measure and then apply some existing clustering methods. However, the edit distance essentially captures the optimal global alignment between a pair of sequences, but ignores many other local alignments that often represent important features shared by a pair of sequences. In addition, the computation of pairwise edit distance can be long and inefficient.

An alternative approach is to utilize the suffix tree structure to explore significant patterns of the protein sequences and use these patterns to evaluate the similarity. More specifically, these patterns can be used to assess the similarity between a pair of protein sequences and between a protein sequence and a cluster (of protein sequences). Under this model, protein sequences belonging to one cluster may subsume to the same (or a similar) probability distribution of amino acids (conditioning on the preceding segment of a certain length), while different clusters may follow different underlying probability distributions. This feature, typically referred to as *short memory*, which is common to many applications, indicates that, for a certain sequence, the empirical probability distribution of the next symbol given the preceding segment can be accurately approximated by observing no more than the last L symbols in that segment. Significant features of such probability distribution can be very powerful in distinguishing different clusters. The suffix tree (or its variation) has been proven to be a very successful model to capture significant patterns in symbol sequences. With regard to the problem addressed in this paper, the suffix tree structure can be utilized to serve as a compact representation to summarize significant patterns shared by members of each cluster.

[1]IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532. E-mail: `ww1@us.ibm.com`

[2]IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532. E-mail: `jiyang@us.ibm.com`

The model can be easily used to identify and organize significant appearances of segments among a cluster of sequences, regardless of the relative positions of these segments within different sequences. These significant patterns imply certain features that is common to a large subset of sequences in the cluster and reveal important statistical properties of the cluster. In practice, they induce a (conditional) probability distribution of the next symbol given the preceding segments, which can be used in estimating the similarity between a sequence and a cluster. The key idea is that, by extracting and maintaining significant patterns characterizing (potential) sequence clusters, one can easily determine whether a sequence should belong to a cluster by calculating the likelihood of (re)producing the sequence under the probability distribution that characterizes the given cluster.

In this paper, we also adopt the suffix tree structure as the foundation to support our similarity measure. In addition, we introduce a novel technique to determine the convergence of each custer and to incorporate available prior knowledge. A pair of upper and lower probabilities is used to quantify the range of the true probability. It has been proven that such probability range is guaranteed to converge to a single value after sufficient trials if the underlying probability distribution remains static. An open gap between the upper and lower probabilities (after sufficient trials) is indeed a good indication of the heterogeneity of the probability distributions, which in turn implies that the corresponding sequences belong to multiple clusters. Therefore, the convergence of the imprecise probabilities can well serve as the touchstone to trigger necessary cluster split and to determine the termination of the clustering process. This provides the opportunity to design an integrated clustering process that allows a more flexible and systematic switch between model training and splitting, comparing to the annealing method, which consists of iterations of soft clustering and progressive refinement.

A suffix tree is maintained for each (potential) cluster of protein sequences in our proposed scheme. Each node x in the suffix tree is associated with an occurrence count $count(x)$ and a vector of n entries, each of which corresponds to a distinct symbol, d, and consists of an empirical probability $prob(d)$ and a pair of lower and upper probabilities $(\underline{prob}(d), \overline{prob}(d))$. The counter $count(x)$ keeps track of the number of occurrences of the path label $\sigma(x)$ of x. The probability vector monitors both the empirical value and the potential range of the conditional probability of observing the symbol, say d, right after $\sigma(x)$ in a protein sequence. Starting from $[0, 1]$ or some other initial range based on the prior knowledge, the range defined by the lower and upper probabilities will be continuously refined and eventually converge during the clustering process.

We experimented our sequence clustering algorithm on a set of 8000 protein sequences of 30 families from the SWISS-PROT database. Within eight minutes, our clustering algorithm finishes the clustering process and achieves a high level of accuracy, e.g., over 93% of sequences are correctly clustered with their family members. Therefore, this clustering algorithm can be very useful in clustering a large set of unlabeled protein sequences.