# A Framework Towards Efficient and Effective Sequence Clustering

Wei Wang
IBM T. J. Watson Research Center
ww1@us.ibm.com

Jiong Yang
IBM T. J. Watson Research Center
jiyang@us.ibm.com

During recent years, analyzing sequence data (particularly in categorical domains) has become increasingly important, partially due to the significant advances in the biological study as well as many other fields. Examples of sequence data include DNA sequences, unfolded protein sequences, text documents, web usage data, system traces, etc.. Previous work on mining sequence data has mainly focused on the frequent pattern discovery. In this project, we focus on the problem of clustering sequence data.

Clustering has been widely recognized as a powerful data mining technique 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. In the context we address, each object is a symbol sequence and all potential features of a sequence are encoded (implicitly) in the specific symbol layout in the sequence. It is interesting to notice that these structural characteristics sometimes can uniquely determine the functional properties of the sequence and often plays a decisive role in meaningful clustering of the sequence.

- A protein sequence is an ordered list of amino acids. The functionality of the protein sequence is solely encoded in the specific ordering of the amino acids. Protein sequences with similar biological functions would share some common signature.
- A text document is essentially a (long) list of alphabet in a certain order and its semantic is uniquely determined by the spelling of each word and the order of words.

Both of the above examples suggest that clustering based on structural characteristics can serve as a powerful tool to discriminate sequences belong to different functional categories. Most research on clustering focused on non-sequential domains and assumed metric space in many cases. In most of these models, the similarity between objects can be defined in a clear and relatively straightforward way. Unfortunately, in the context we address here, each object is a sequence of symbols and there is no obvious similarity measure that can be applied directly. This is in part because (1) each sequence is composed of symbols of non-numerical nature; (2) the length of a sequence can easily run up to thousands or even beyond; and (3) different sequences may be of vastly different lengths. It is not feasible to map a sequence into a point in a high dimensional space and then utilize existing clustering algorithms developed on metric space.

To tackle this problem, we propose a novel model for sequence cluster by exploring significant patterns of sequence formation. Significant patterns charactering sequence behaviors can be utilized to assess the similarity between a pair of sequences and between a sequence and a cluster (of sequences). In order to achieve this, an underlying model needs to be assumed to formalize the set of potential features. Sequences belonging to one cluster may subsume to the same probability distribution of symbols (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 structure is utilized to serve as the compact representation to summarize significant patterns shared by members of each cluster. 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. A randomized algorithm is designed to cluster the set of sequences into a number of $k$ clusters where $k$ can be a user specified parameter. Comparing to alternative similarity measures based on edit distance and hidden Markov model, the proposed scheme can render two orders of magnitude performance improvement while producing results of comparable quality to that of hidden Markov model and of even better quality than that of edit distance. Last but not least, this approach can also be utilized to efficiently identify outliers in a sequence database.