Read the CLUSEQ paper from the lecture. Answer the following question: What is the difference between edit distance, Hidden Markov Models and Probabilistic Suffix Trees? What is the advantage of PST w.r.t the other models?

Solution:

In order to analyze the categorical sequences for finding similarity between objects containing similar features , several methodological approaches are available. It is used in various fields like protein sequencing (biology) and text documents for data mining. This reveals unknown object groups/categories that may leading to a better understanding. The "sequential" characteristics play a crucial role in determining the properties of a sequence thereby contributing to clustering.

Edit Distance is one of the methods to quantify the dissimilarity by counting the minimum number of operations required to transform one string into the other. Though it has several applications, it is inefficient as it ignores the local alignment contributing to its global nature alignment. These overlooked features play a critical role in forming clusters. The blocked operations which may alleviate this weakness, is NP hard.The q gram based method which uses generic symbol sequences to compute similarity, the quality of the clusters generated is not effective due to ignorance of sequential relationship (ordering, correlation, dependency, etc.) among them.

HMM(**hidden Markov model)** , a dynamic Bayesian Network is another methodological approach which relies on the hypothesis that the observed data is generated by modeling the underlying process. It states that the conditional probability of a state in a sequence depends on a fixed number of past states, which is defined by the order (memory) L of the chain. The number of parameters to be estimated increases exponentially with L; thus, fitting high-order models becomes infeasible.

Suffix Tree a compressed trie containing all the suffixes of the given text as their keys and positions in the text as their values. To support efficient maintenance and retrieval of the probability entries, probabilistic suffix tree **(PST)**, is employed efficiently organize statistical properties of a sequence cluster. It is evidently proved to be a successful approach which is utilized to serve as the compact indexing structure to organize significant segments and associated conditional probability entries of each cluster. This enables similarity estimation to be performed very fast, offering many advantages over alternative methods and plays a dominant role in increasing the overall performance of the CLUSEQ clustering algorithm ,as discussed in the paper.The probability to predict a sequence is :

$$P_S(\sigma) = P_S(s_1) \times P_S(s_2|s_1) \times \ldots \times P_S(s_l|s_1 \ldots s_{l-1})$$

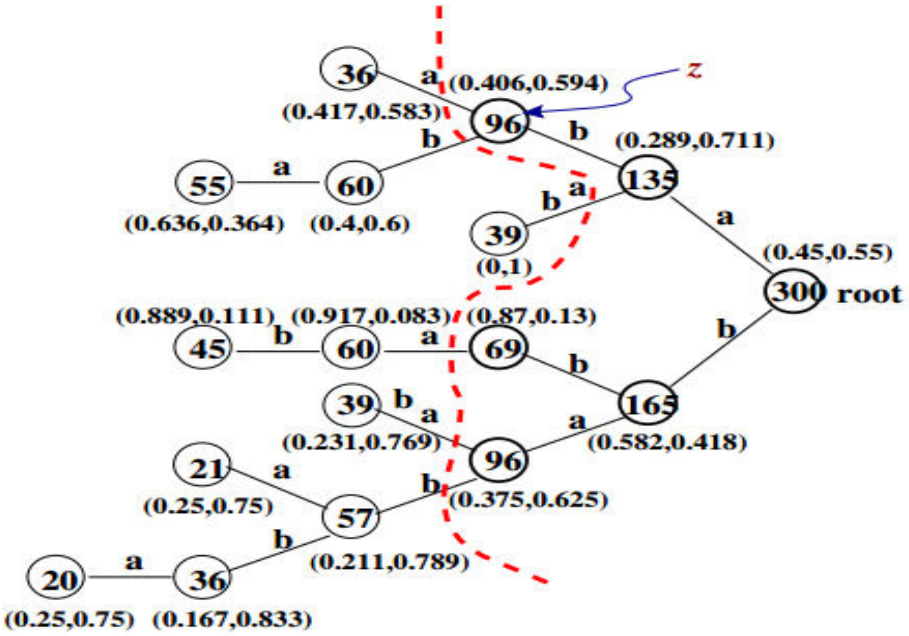

Figure 1. A Probabilistic Suffix Tree

The PST computes the similarity of the longest significant suffix , it first builds a suffix tree by building the suffix tree on the reversed sequences (instead of the original sequences).It associates a Count C and computes its significance to find the prediction node by traversing the tree .The entry retrieved by node is used as the estimation value of the sequence .

The CLUSEQ algorithm , designed to cluster sequences into a set of possibly overlapped clusters, where the number of clusters and the amount of outliers can be automatically adjusted during the clustering process,uses a probabilistic suffix tree to store significant features (i.e., CPD) of each sequence cluster. It uses as input optimal k(the number of clusters), t(similarity threshold) and c(count of the length of the segment)

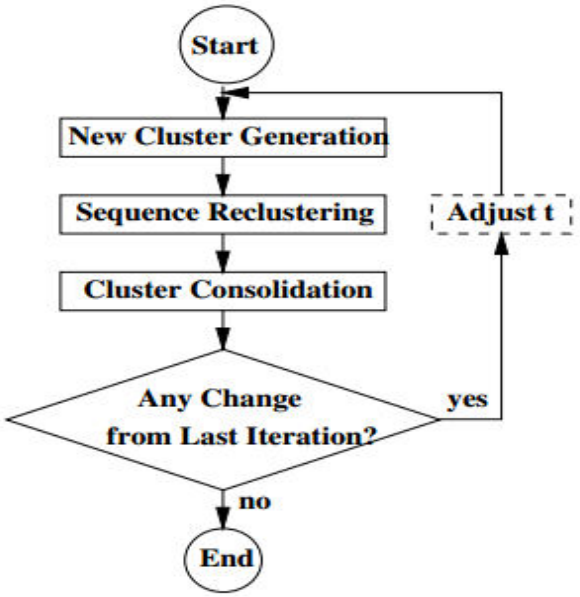

Figure 2. The Flowchart of the CLUSEQ Algorithm

ADVANTAGES:

The advantages of PST as compared to HMM and Edit distance is as follows:

- It uses adjusted probabilities instead of raw the raw empirical probabilities in the computation so that no symbol will have an absolutely zero probability to appear no matter what segment is observed before it. Also, it can be performed on the fly during the similarity estimation.

- The accuracy of the algorithm is better than HMM,Edit Distance(ED), edit distance with block operation (EDBO),and q-gram.

- It automatically changes the value of t(the similarity threshold) from the initial setting after a particular iterations and consolidating the clusters to prevent insignificant clusters to be present with overlapping features.

- The accuracy of CLUSEQ is immune to the increase of outliers.

- The response time and the complexity of PST is better as compared to HMM and ED. As shown in the paper,the performance with high accuracy and better response rate is showcased.

Table 2. Model Comparison

Model	CLUSEQ	ED	EDBO	HMM	q-gram
Percentage of Correctly Labeled Proteins	82%	23%	80%	81%	75%
Response Time (sec.)	144	487	13754	3117	132

- The CLUSEQ algorithm is very robust in the face of erroneous initial parameter setting.

REFERENCES:

1) Jiong Yang and Wei Wang. CLUSEQ: Efficient and Effective Sequence Clustering.
2)N. Slonim and N. Tishby. The power of word clusters for text classification. Proc. of ECIR, 2001