# A Generic Framework for Efficient and Effective Subsequence Retrieval

Haohan Zhu, George Kollios, Vassilis Athitsos

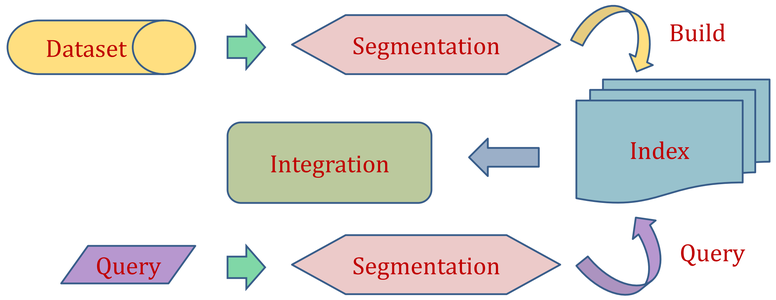
*Abstract*: We propose a general framework for matching similar subsequences in both time series and string databases. The matching results are pairs of query subsequences and database subsequences. The framework finds all possible pairs of similar subsequences if the distance measure satisfies the "consistency" property, which is a novel property introduced by our framework. We show that most popular distance functions, such as the Euclidean distance, DTW, ERP, the Frechet distance for time series, and the Hamming distance and Levenshtein distance for strings, are all "consistent". We also propose an index structure for metric spaces named "reference net". The reference net is an unsupervised index which costs O(n) space, where n is the size of the dataset. The experiments demonstrate the ability of our method to improve retrieval performance when combined with diverse distance measures. The experiments also illustrate that the reference net has a better running time than cover trees and the maximum variance method, while all three methods have similar costs in terms of space.

## Consistency

We call distance δ a "consistent" distance measure if it obeys the following property: if Q and X are two sequences, then for every subsequence SX of X there exists a subsequence SQ of Q such that $\delta(SQ, SX) \leqq \delta(Q, X)$.

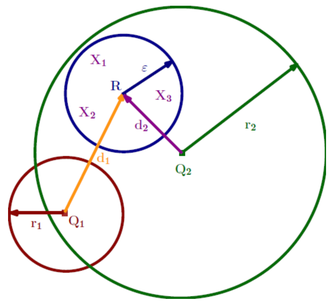
We can segment X by non-overlapping fixed length time windows and segment Q by sliding windows. If SX and SQ are similar subsequences with lengths larger than λ, there must be a pair of time segments that are similar with lengths of λ/2.

## Framework

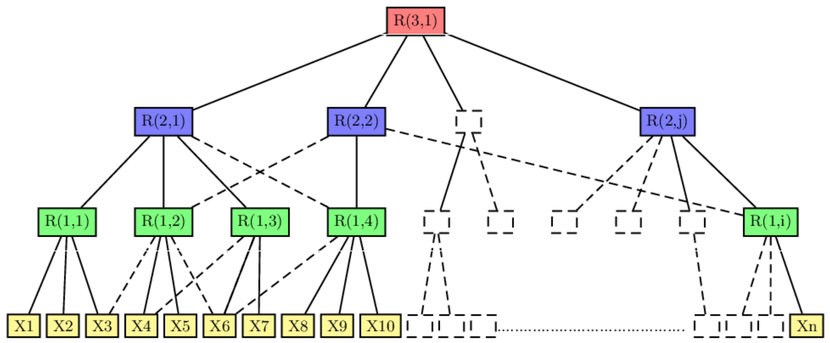



## Reference-based Method

As shown in the following figure, since $\delta(Q_2, R) + \varepsilon < r_2$, we can claim that all distances between $Q_2$ and every $X_i$ are smaller than $r_2$. Also because $\delta(Q_1, R) - \varepsilon > r_1$, we can claim that all distances between $Q_1$ and every $X_i$ are larger than $r_1$.

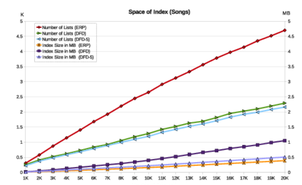


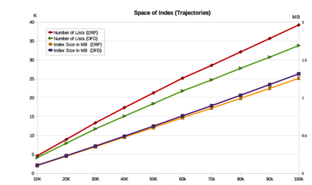

## Reference Net Structure

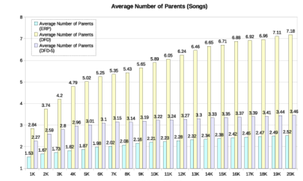


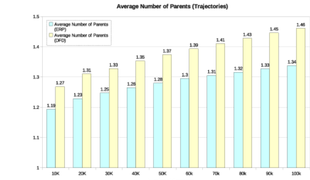

## Index

Reference Net has linear space O(n)



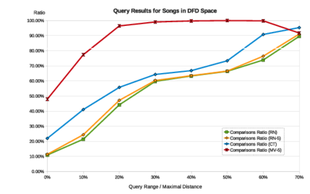



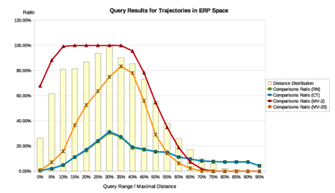



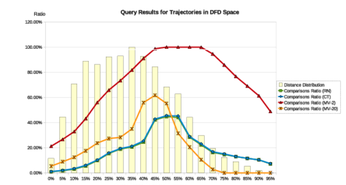




## Query

Reference Net is better than Cover Tree and Maximum Variance


## Applications