Time Series Analysis with Applications in Bioinformatics

Coleman Yu

Abstract

Time series data are ubiquitous across scientific disciplines, necessitating robust computational techniques for analysis and retrieval. This thesis addresses two critical challenges in the domain of time series data mining with applications in bioinformatics: the development of more expressive distance measures, with the usage of similarity search, and the novel application of time series classification to solve complex biological problems.

In the first part, we address the limitations of existing similarity measures. Similarity search is a core subroutine in tasks such as classification and motif finding. While Dynamic Time Warping (DTW) and Uniform Scaling (US) are prevailing measures for handling local distortions and global scaling, respectively, and some studies have demonstrated that combining both DTW and US is necessary to obtain meaningful results, the current approaches are limited to applying a single scaling factor to the entire sequence before applying DTW. We argue that since distinct phases of a process often evolve at different speeds, a single scaling factor is insufficient. We introduce the first distance measure that achieves invariance to multiple scaling factors. We also provide lower bounding techniques to facilitate efficient computation of the proposed distance measure. This method better reflects the similarity between time series with multiple phases and provides a clearer understanding of the data.

In the second part, we demonstrate the applied power of time series analysis within the field of bioinformatics. Bioinformatics operates at the intersection of Biology, Biotechnology, and Informatics. In this work, we formulate a specific

Biology problem, which is predicting Human Dicer Cleavage sites in microRNA biogenesis, into a machine learning framework. Due to the current limitations of Biotechnology, we are constrained to utilizing 1-D RNA sequence inputs rather than fully 2-D data; the latter is more expensive to obtain. We propose MTSC-Cleav, a method that encodes RNA sequences and the probabilities of base pairs in predicted secondary structures into time series data. By doing this, we frame the problem of predicting Human Dicer Cleavage sites into a Multivariate Time Series Classification (MTSC) problem. Unlike existing approaches that rely on opaque deep neural networks or complex feature engineering, our approach is simple and computationally efficient. Experiments demonstrate that MTSCCleav achieves comparable accuracy to state-of-the-art methods while delivering a 3.7X to 28.8X speedup. Furthermore, our perturbation experiments reveal that regions near the center of pre-miRNAs are essential for cleavage site prediction.

Collectively, this thesis advances the fields of time series data mining and bioinformatics by proposing a new, more expressive distance measure and demonstrating the use of time series analysis to address fundamental biological questions.

Publications

This thesis is based on the following papers.

- (Chapter 3) **Coleman Yu**, Tatsuya Akutsu, Raymond Chi-Wing Wong. Scaling with Multiple Scaling Factors and Dynamic Time Warping in Time Series Searching. Preparation for the submission of *IEEE Access*.

- (Chapter 4) **Coleman Yu**, Raymond Chi-Wing Wong, Tatsuya Akutsu. MTSCCleav: a Multivariate Time Series Classification (MTSC)-based Method for Predicting Human Dicer Cleavage Sites. Submitted to *BMC Bioinformatics* on 14 July 2025, under the first review.

Contents