Multidimensional Time Series Analysis for Protein Structure Classification

Michaela Areti Zervou
University of Crete and FORTH-ICS, zervou@ics.forth.gr

Effrosyni Doutsi
FORTH-ICS, edoutsi@ics.forth.gr

Pavlos Pavlidis
FORTH-ICS, pavlidis@ics.forth.gr

Panagiotis Tsakalides
University of Crete and FORTH-ICS, tsakalid@ics.forth.gr

ABSTRACT

In the last decades, many studies have explored the potential of utilizing time series analysis tools such as recurrence quantification analysis (RQA), horizontal visibility graphs (HVG) and others to solve one of the most significant problems in bioinformatics, the structural class prediction of a protein. In this work we propose novel architectures based on the aforementioned algorithms, demonstrating the superiority of the best-performed scheme when compared against the state-of-the-art on the classification of real protein sequences.

1 INTRODUCTION

The rapid progress of genomics over the last few decades has resulted in a massive amount of protein sequence. To this end, a growing need for effective protein structure classification architectures has emerged. Numerous studies [4, 5] demonstrate the potential of transforming the amino acid sequence into a time series and then utilizing powerful time series analysis techniques such as Recurrence Quantification Analysis (RQA) [1], Horizontal Visibility Graphs (HVG) [2] or a combination of both, for extracting meaningful information from the protein sequence data that can be presented as a two-dimensional time series where each dimension is processed independently. However, the enhanced performance of those techniques comes at the expense of time and memory complexity as well as with an inefficient processing of the multidimensional data.

To overcome these limitations, HVG for multidimensional data (mdHVG) [5] and the Generalized multidimensional Recurrence Quantification Analysis (GmdRQA) [3] are employed to operate directly on the multidimensional data, and data-driven estimation schemes of the RQA/GmdRQA parameters are proposed and described in Section 2.

2 PROPOSED ARCHITECTURES

In this work, we designed novel data-driven protein structural class prediction architectures using RQA, GmdRQA, and a combination of HVG and mdHVG with the two aforementioned techniques as depicted in Fig. 1. RQA and GmdRQA are parametric models, thus their parameter selection should be done effectively. Therefore, we perform a parameter selection scheme based on (i) Grid Search (GS) that is mainly utilized in the literature, and propose and conduct two more case studies: (ii) the RQA/GmdRQA parameters are found for each protein in a Data-Driven (DD) fashion, and (iii) the Most Frequent (MF) values are computed as the product of a statistical analysis of the optimal set of parameters. These three parameter selection approaches are employed by RQA and GmdRQA independently, resulting in 6 different algorithms. Thereafter, HVG and mdHVG are also combined with each one of the 6 aforementioned RQA-based algorithms and several quantification measures are computed for both RQA-based and HVG-based architectures. Later Fisher's Linear Discriminant Analysis (FDA) classifier is applied on the each resulted feature matrix for each study case. Finally, every proposed architecture is evaluated in terms of classification accuracy, feature multitude and computational complexity.

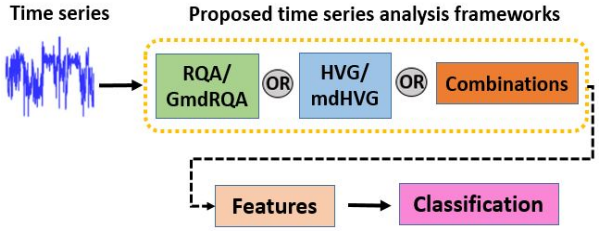

Figure 1: General pipeline of the proposed classification architecture.

The study on real protein data revealed the superiority of mdHVG scheme, when compared to the rest proposed approaches and the state-of-the-art, in terms of classification accuracy, feature multitude and running time complexity.

3 CONCLUSIONS

The best-performed scheme, mdHVG, can be seen as a profitable protein structure classification architecture as it outperforms the state-of-the-art architectures in both accuracy and computational efficiency. A graph convolutional neural network architecture based on the graph representation of the proteins produced by the mdHVG method will be considered as an extension of this study.

This research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under HFRI faculty grant no. 1725, and by the Stavros Niarchos Foundation within the framework of the project ARCHERS. .

REFERENCES

[1] Eckmann J.P. et al. 1987. Recurrence plots of dynamical systems. *Europh. Lett.* 4, 9 (1987), 973.
[2] Lacasa L. et al. 2008. From time series to complex networks: The visibility graph. *Proc. of the National Academy of Sciences* 105, 13 (2008), 4972–4975.
[3] M.A. Zervou et al. 2019. Automated Screening of Dyslexia via Dynamical Recurrence Analysis of Wearable Sensor Data. In *2019 IEEE 19th International Conf. on Bioinformatics and Bioengineering (BIBE)*. IEEE, 770–774.
[4] M.A. Zervou et al. 2021. Structural classification of proteins based on the computationally efficient recurrence quantification analysis and horizontal visibility graphs. *bioRxiv* (2021), 2020–10.
[5] M.A. Zervou et al. 2021. Visibility Graph Network of Multidimensional Time-Series Data for Protein Structure Classification. In *29th European Signal Proc. Conf. (EUSIPCO)*. IEEE.