# GrammarViz 2.0: a tool for grammar-based pattern discovery in time series

Pavel Senin[1], Jessica Lin[2], Xing Wang[2], Tim Oates[3], Sunil Gandhi[3],
Arnold P. Boedihardjo[4], Crystal Chen[4], Susan Frankenstein[4], and Manfred Lerner[5]

[1] University of Hawaii, Manoa, ICS Dept., CSDL, senin@hawaii.edu
[2] George Mason University, Dept. of Computer Science,
{jessica, xwang24}@gmu.edu
[3] University of Maryland, Baltimore County, Dept. of Computer Science,
oates@cs.umbc.edu, sunilga1@umbc.edu
[4] U.S. Army Corps of Engineers, Engineer Research and Development Center,
{arnold.p.boedihardjo, crystal.chen,
susan.frankenstein}@usace.army.mil
[5] SAP Germany, manfred.lerner@sap.com

**Abstract.** The problem of frequent and anomalous patterns discovery in time series has received a lot of attention in the past decade. Addressing the common limitation of existing techniques, which require a pattern length to be known in advance, we recently proposed grammar-based algorithms for efficient discovery of variable length frequent and rare patterns. In this paper we present GrammarViz 2.0, an interactive tool that, based on our previous work, implements algorithms for grammar-driven mining and visualization of variable length time series patterns.

## 1 Introduction

The ability to efficiently detect frequent and anomalous patterns in time series allows for the exploration, summation, and compression of data. In addition, such information is crucial to a variety of application domains where these patterns convey critical and actionable information, such as health care, equipment safety, and security. Furthermore, these patterns are often used as input features for data mining tasks, such as association rule mining and classification.

Previously, we defined time series motifs (frequent patterns) [1] and time series discords (anomalous patterns) [2], and proposed efficient *exact* solutions for their discovery based on Symbolic Aggregate Approximation (SAX) [3]. While there has been a great amount of follow-up work on the discovery of both pattern types [4], one common limitation of currently available techniques is that they require the length of a potential motif or discord to be specified as input. This is unreasonable for most real-world problems as such information may not be known in advance, and patterns of different lengths may co-exist in the data.

This research is partially supported by the National Science Foundation under Grant No. 1218325 and 1218318.

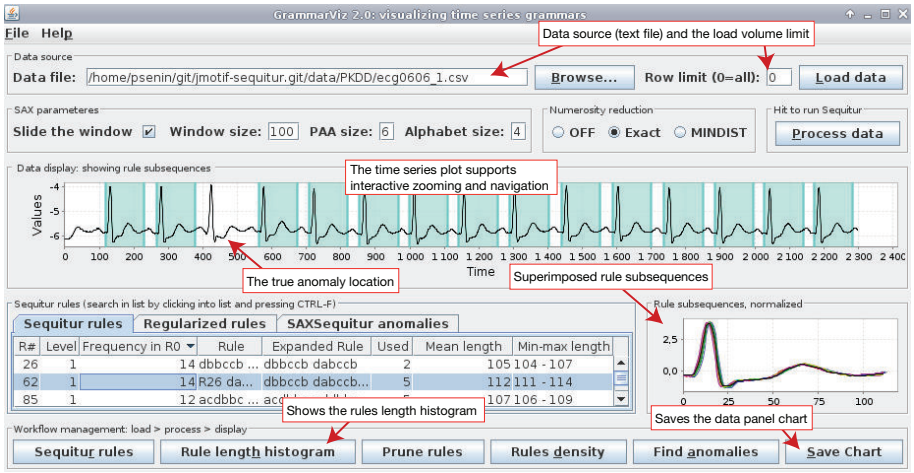


Fig. 1: An example of a recurrent grammar rule (i.e. ***motif***) discovery in the ECG dataset using GrammarViz 2.0. Note, that the highlighted motif does not cover an anomalous heartbeat and that rule-corresponding subsequences vary in length.

Addressing this limitation, we recently proposed an alternative solution for the discovery of variable-length motifs [5] and anomalies [6] based on SAX discretization and the Sequitur grammar inference algorithm [7]. We showed that our algorithm is able to efficiently discover *co-existing variable-length approximate motifs and anomalies* without any prior knowledge about their length, shape, or minimal occurrence frequency. In this work, we present a time series pattern discovery application called GrammarViz 2.0 that can simultaneously discover variable-length motifs and anomalies.

## 2 Our approach and the tool for time series patterns mining

Our approach is built on a three phase process: time series discretization, context free grammar induction, and motif/anomaly detection. The first step is to model the time series as discretized elements and convert it into a symbolic representation. The second step is to parse the symbolic series and decompose it into a context free grammar [5, 6]. Since rules of a context free grammar are hierarchically organized, it is possible to establish the probability of occurrence of a time series subsequence using its corresponding rule hierarchy and rule counts in the entire time series. Intuitively, since each grammar rule represents a discretized subsequence *pattern* of the input time series, frequently used rules are likely to correspond to recurrent subsequences, while infrequently used rules are likely to correspond to rare subsequences.

Next, we discuss the detailed steps of the above approach and its implementation in our grammar-driven workflow for time series patterns discovery.

### 2.1 Dimensionality reduction and discretization with SAX

Time series are real-valued data whereas grammar induction algorithms are designed for discrete values. We rely on SAX [3] to discretize the input time series. For time se-

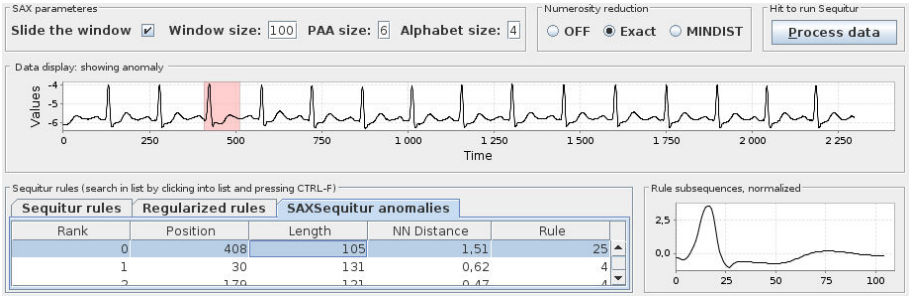


Fig. 2: An example of an anomalous grammar rule discovered in the ECG dataset which corresponds to a very subtle anomaly in the ST wave annotated by an expert [2].

ries $T$ of length $m$, SAX obtains a lower-dimensional representation by first performing a $z$-normalization then dividing the time series into $w$ equal-sized segments. Next, for each segment, SAX computes a mean value and maps it to a symbol according to a predefined set of breakpoints dividing the data space into $\alpha$ equiprobable regions, where $\alpha$ is the user specified alphabet size. While dimensionality reduction is a desirable feature for exploring global patterns, the high compression ratio ($m/w$) significantly affects performance in cases where localized phenomena are of interest. Thus, for the local pattern discovery, and specifically for motif and anomaly detection, SAX is typically applied to a set of subsequences that represent local features – a technique called subsequence discretization [1] which is implemented via a sliding window.

Our tool implements both global and local discretization and allows an interactive tuning of discretization parameters using "SAX parameters" panel (Fig.1). In addition, next to the SAX parameters selection, users can toggle the numerosity reduction strategy, which not only mitigates for trivial and degenerate pattern discovery [2, 3], but enables an essential feature of our technique – the discovery of variable-length co-existing patterns [5, 6].

### 2.2 Context free grammar induction with Sequitur

For grammar inference, we rely on Sequitur - a linear time and space algorithm that derives a context-free grammar from a string incrementally [7]. By identifying frequent subsequences in the input string, the algorithm builds a compact context-free grammar reflecting the input string specificity. In addition, we are currently extending our application with mSequitur algorithm implementation that introduces a merging operator and is capable of further grammar reduction by generalization [11].

Since Sequitur requires no input parameters, in a single "Process data" step (Fig.1) our tool performs both discretization and grammar induction procedures. Once grammar is built, its rules are presented to the user in a table format enabling efficient examination and exploration of rules and their corresponding subsequences. GrammarViz 2.0 shows rule locations on the original time series and superimposes all rule subsequences on a separate panel. This allows visual evaluation of the results from selected parameters as well as their interactive tuning (Fig.1).

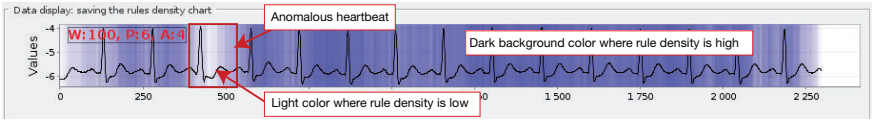


Fig. 3: An example of the "Data display" panel showing the "Rule density" plot used for highly efficient approximate anomaly discovery through visual examination.

### 2.3 Exploiting context-free grammar for pattern discovery

***Motif discovery***. With the capability to sort the rule table by the rule usage frequency, as well as the effective visual presentation of grammar rules, GrammarViz 2.0 allows user to navigate the rules and visually inspect their corresponding subsequences ("motifs").
***Discord discovery***. GrammarViz 2.0 enables anomaly detection in two ways: by integrating grammar induction in the HOTSAX discord discovery framework [2] (Fig.2), and by visualization of the grammar rule density (Fig.3). Both approaches allow the user to visually evaluate potential anomalous rules and their corresponding subsequences.

## 3 Target Audience and Similar Applications

As time series are often used as a proxy to represent a large variety of wide ranging real-life phenomena, the GrammarViz 2.0 application targets diverse audiences including researchers, practitioners, engineers, medical specialists, and safety and security personnel. While other time series pattern visualization tools exist [9, 10], we are not aware of any tool that has the same capabilities as GrammarViz 2.0; namely, the discovery of hierarchical patterns and variable-length motifs and discords.

## References

1. Lin, J., Keogh, E., Patel, P., and Lonardi, S.: Finding Motifs in Time Series. The 2nd Workshop on Temporal Data Mining, the 8th ACM Int'l Conference on KDD. 53–68, (2002)
2. Keogh, E., Lin, J., Fu, A.: HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. In Proc. ICDM. 226–233, (2005)
3. Patel, P., Keogh, E., Lin, J., Lonardi, S.: Mining Motifs in Massive Time Series Databases. In Proc. ICDM, (2002)
4. Chandola, V., Cheboli, D., and Kumar, V.: Detecting Anomalies in a Time Series Database. CS Technical Report 09–004, (2009)
5. Li, Y., Lin, J., and Oates, T.: Visualizing variable-length time series motifs. In Proc. of the 2012 SIAM International Conference on Data Mining, 895-906, (2012)
6. Senin, P., Lin, J., Wang, X. Oates, T., Boedihardjo, A.P., Chen, C., Frankenstein, S., Gandhi, S.: Grammar-driven anomaly discovery in time series. CSDL Techreport 14-05, (2014)
7. Nevill-Manning, C. and Witten, I.: Identifying Hierarchical Structure in Sequences: A linear-time algorithm. Journal of Artificial Intelligence Research, 7, 67-82, (1997)
8. Paper authors. Supporting webpage: http://github.com/GrammarViz2
9. Lin, J., Keogh, E., Lonardi, S., Lankford, J., Nystrom, D.: Visually mining and monitoring massive time series. In Proc. 10th ACM SIGKDD Intl. Conf. on KDD, 460-469 (2004)
10. Hao, M., Marwah, M., Janetzko, H., Dayal, U., Keim, D., Patnaik, D., Ramakrishnan, N. and Sharma, R. K.: Visual Exploration of Frequent Patterns in Multivariate Time Series. Information Visualization, Vol. 11, No. 1, 71-83 (2012)
11. Oates, T., Boedihardjo, A., Lin, J., Chen, C., Frankenstein, S., Gandhi, S.: Motif discovery in spatial trajectories using grammar inference. In Proc. of ACM Intl. Conf. on Information and Knowledge Management (CIKM) (2013)