# Interactive motif discovery in time series with persistent homology

Thibaut Germain[1]✉, Charles Truong[1], and Laurent Oudre[1]

Université Paris Saclay,Université Paris Cité, ENS Paris Saclay,CNRS, SSA, INSERM, Centre Borelli, F-91190, Gif-sur-Yvette, France.
{firstname.lastname}@ens-paris-saclay.fr

**Abstract.** Time series analysis based on recurrent patterns, also called motifs, has emerged as a powerful approach in various domains. However, uncovering recurrent patterns poses challenges and usually requires expert knowledge. This paper introduces an interactive version of the PersistentPattern algorithm (PEPA), which addresses these challenges by leveraging topological data analysis. PEPA provides a visually intuitive representation of time series, facilitating motif selection without needing expert knowledge. Our work aims to empower data mining and machine learning researchers seeking deeper insights into time series. We provide an overview of the PEPA algorithm and detail its interactive version, concluding with a demonstration of abnormal heartbeat detection.

**Keywords:** Time series · Motif discovery · Persistent homology.

## 1 Introduction

Time series representations, based on their recurrent patterns, have found success across various machine learning tasks and domains. Application examples include forecasting energy consumption [3] or stock trends [6], identifying abnormal heartbeats [1], and uncovering physiological behaviors [2]. Pattern-based representations enhance scalability and generalization of machine learning algorithms while retaining insightful information about the time series.

However, the initial task of retrieving recurrent patterns and their occurrences poses challenges [7]. Known as motif discovery, several algorithms have been proposed to solve this task [7], many of which exhibit quadratic time complexity in the length of the time series. These algorithms typically rely on three parameters: the number of motifs to discover, the length of motifs, and a similarity threshold between motif occurrences. Yet, setting these parameters without expert knowledge often involves time-consuming trial-and-error strategies [7].

In a recent work [5], we introduced the PersistentPattern algorithm (PEPA) which discovers motifs of variable lengths. PEPA addresses the limitations of prior algorithms by postponing the step of setting the number of motifs and the similarity threshold. The algorithm first summarizes a time series with a persistence diagram, a key data representation in topological data analysis. This diagram offers an intuitive and visual summary of time series, facilitating motif

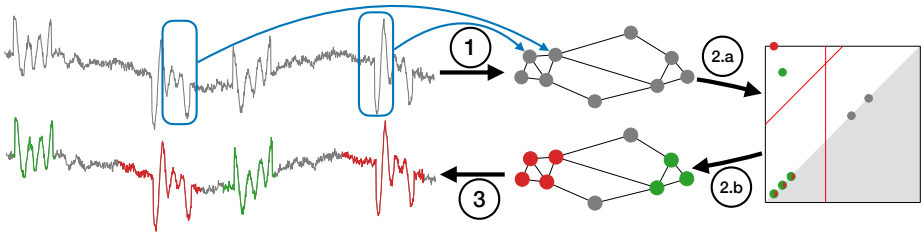


**Fig. 1.** Workflow of the PersistentPattern algorithm

selection. Furthermore, PEPA effectively retrieves motifs from the persistence diagram using an algorithm with linear time complexity.

Thanks to its visual interpretation and efficiency, we present an interactive version of PEPA algorithm, enabling extensive exploration of time series recurrent patterns. Our target audience includes data mining and machine learning researchers seeking to deepen their understanding of the structural properties of some time series or to collect recurrent patterns.

In contrast, a prior system [8] relied on a grammar induction algorithm, necessitating a symbolic time series. However, motif discovery sensitivity to this discretization [5] demands expert tuning.

In what follows we briefly present PEPA algorithm and detail the operating system of its interactive version. We conclude with a system demonstration showcasing the detection of abnormal heartbeats from an electrocardiogram (ECG).

## 2 System overview

**PersistentPattern (PEPA):** The algorithm relies on a graph to encode the structural relationships between all subsequences of a time series. It also uses persistent homology to identify and isolate motifs. PEPA can be broken down into three steps illustrated in Figure 1:

1. **From time series to graph, Fig.1-1:** Transforms a time series into a graph where nodes are subsequences and edges are weighted with a distance between subsequences. The distance function $d_\epsilon : \mathbb{R}^l \times \mathbb{R}^l \mapsto [0, 2]$ is the Euclidean distance between ($\epsilon \in \{LT, Z\}$)-normalized subsequences [4].
2. **Graph clustering with persistent homology, Fig.1-2.a-b:** Identifies clusters representing motifs from the persistence diagram and separates them from irrelevant parts of the time series with two thresholds (red lines).
3. **From clusters to motif sets, Fig.1-3:** Merges temporally adjacent subsequences in each cluster to form the variable length motifs.

**Interactivity & Interpretability:** The persistence diagram interprets the structure of the time series: motifs are represented by points in the top-left corner, while irrelevant parts of the series are situated on the right, and motif subdivisions are located in the lower-left corner. Motifs are efficiently retrieved


**Fig. 2.** Interface, the user has discovered normal and abnormal heartbeats in an ECG.

from the persistence diagram (steps 2.b & 3) when motif points are isolated with a vertical threshold and an off-diagonal one. However, motifs can be sensitive to these thresholds. The threshold adjustment can be facilitated without additional computation cost by emphasizing the visual interpretation of the persistence diagram. To that end, it suffices to modify the distance function between subsequences carefully. We introduce the $(\alpha, \beta)$-rectified distance:

$$d(x, y) = 2\sqrt{f_{\alpha,\beta}(d_\epsilon(x, y))/f_{\alpha,\beta}(2)}$$

where $(x, y) \in \mathbb{R}^l \times \mathbb{R}^l$ are subsequences, $f_{\alpha,\beta}(x) = \tanh(\alpha\beta^2) + \tanh(\alpha(x^2 - \beta^2))$ with $\alpha \in \mathbb{R}_+^*$ and $\beta \in [0, 2]$. Intuitively, $d$ polarizes points of the persistence diagrams by controlling the variability between subsequences through the distance slope $(\alpha)$ and its offset at the origin $(\beta)$.

**Operating system:** The system is implemented in Python with the Dash library. It is accessible from a webpage[1] or can be run locally[2]. Following PEPA's workflow, the user interface is divided into three blocks, Figure 2:

- **Upper block (red):** Associated with step 1, the user uploads a time series, sets parameters related to the graph construction, and runs it.
- **Middle left block (green):** Associated with step 2, it is the core interactive component of the system. The user can modify the distance function and set the threshold from the resulting persistence diagram.
- **Middle right & lower blocks (blue):** Associated with step 3, the lower block displays the time series and highlights the discovered motifs. The middle-right block displays motifs individually.

After step 1, the system stores the time series, the graph, and the persistence diagram. These elements allow smooth back and forth between steps 2 and 3,

[1] Webpage: https://persistent-pattern-discovery.onrender.com
[2] Github: https://github.com/thibaut-germain/Persistent_Pattern_Discovery_App

providing a playground for deepening the user's knowledge about the time series. We also provide detailed guidelines and information in the system itself.

## 3 Use case: Detection of abnormal heartbeats in an ECG

The dataset `MIT-BIH`[3] compiles ECGs from patients experiencing premature ventricular contractions (PVCs). In Figure 2, we explore a 16-second segment of a recording. Here, the distance parameters are automatically adjusted using the `Spread` button to evenly distribute points across the x-axis in the persistence diagram. The persistence diagram suggests two motifs and a clear distinction between motifs and irrelevant sections. By clicking on the `Automatic cut` button, the user ran a heuristic to set both thresholds [5]. The signal and the displayed motifs reveal that the red motif corresponds to normal heartbeats, while the blue motif corresponds to abnormal ones accounting for PVCs. A "normal" heartbeat has not been discovered; its persistent representation corresponds to the closest points at the right of the vertical threshold. The system's responsiveness enables quick threshold adjustment for accurate discovery. We also provide a demonstration video[4].

**Acknowledgments.** This work was supported by grants from Région Ile-de-France (DIM MathInnov). Charles Truong is funded by the PhLAMES chair of ENS Paris-Saclay.

## References

1. Elangovan, R., Padmavathi, S.: A review on time series motif discovery techniques an application to ecg signal classification: Ecg signal classification using time series motif discovery techniques. International Journal of Artificial Intelligence and Machine Learning **9**(2), 39–56 (2019)
2. Feng, T., Narayanan, S.S.: Discovering optimal variable-length time series motifs in large-scale wearable recordings of human bio-behavioral signals. In: International Conference on Acoustics, Speech and Signal Processing. pp. 7615–7619. IEEE (2019)
3. Funde, N.A., Dhabu, M.M., Paramasivam, A., Deshpande, P.S.: Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data. Sustainable cities and society **46**, 101415 (2019)
4. Germain, T., Truong, C., Oudre, L.: Linear-trend normalization for multivariate subsequence similarity search. In proceedings of the international conference on data engineering workshops (ICDEW), Utrecht, Netherlands (2024)
5. Germain, T., Truong, C., Oudre, L.: Persistence-based motif discovery in time series, submitted to ieee transactions on knowledge and data engineering. http://www.laurentoudre.fr/publis/TKDE2024 (2024)
6. Huang, Y., Mao, X., Deng, Y.: Natural visibility encoding for time series and its application in stock trend prediction. Knowledge-Based Systems **232**, 107478 (2021)

[3] https://physionet.org/content/svdb/1.0.0/
[4] Demonstration video: https://youtu.be/F2bwCKiR-i8

7. Schäfer, P., Leser, U.: Motiflets: Simple and accurate detection of motifs in time series. Proceedings of the VLDB Endowment **16**(4), 725–737 (2022)
8. Senin, P., Lin, J., Wang, X., Oates, T., Gandhi, S., Boedihardjo, A.P., Chen, C., Frankenstein, S., Lerner, M.: Grammarviz 2.0: a tool for grammar-based pattern discovery in time series. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part III 14. pp. 468–472. Springer (2014)