Efficiently Mining Frequent Representative Motifs in Large Collections of Time Series

Stijn J. Rotman[1], Boris Cule[1], and Len Feremans[2]

[1] Department of CSAI, Tilburg University
[2] Department of Computer Science, University of Antwerp
s.j.rotman@tilburguniversity.edu

Keywords: Time series, Motif discovery, Pattern mining

An important task in time series data mining is motif discovery. Motifs are repeated patterns in time series. Most motif discovery algorithms focus on finding motifs within single time series [7], the most prolific method for discovery being the Matrix Profile, which finds motif pairs [8,11,10,5]. However, the Ostinato algorithm [2] allows the discovery of *consensus motifs*, motifs that occur frequently *across* time series. As is the case for the Matrix Profile, Ostinato requires a user-defined motif length. Though VACOMI [9] makes enumeration over several lengths more efficient, variable-length consensus motif discovery remains a costly procedure.

We introduce FRM-Miner, which can find *frequent* motifs of variable length in time series databases. In contrast to consensus motifs, frequent motifs do not necessarily occur in every time series across a database, which is a constraint that is more typical for real-world scenarios. Instead, the optimal support and length are determined automatically for each motif. Motifs of varying length are of particular interest to us as they are often encountered in real-world applications [9,1]. In order to illustrate our main contributions, the main steps of FRM-Miner are visualised in Fig. 1, using contours of cattle in the MPEG-7 data set [3]. This algorithm is much more efficient than Ostinato and VACOMI. This encore abstract provides a summary of prior work by the same authors that was presented at IEEE BigData'23 [6].
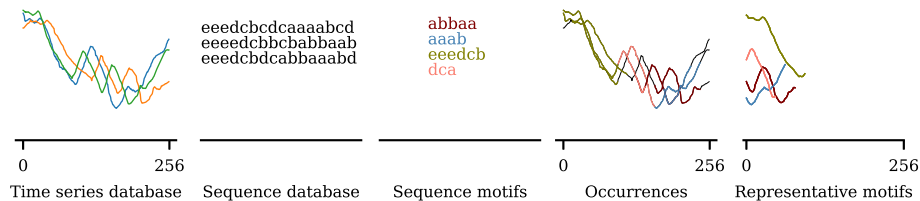
Fig. 1. Pipeline of all steps in FRM-Miner. The time series are discretised to sequences. Frequent sequence motifs are discovered and mapped back to time series occurrences, which are then used to construct representative motifs.
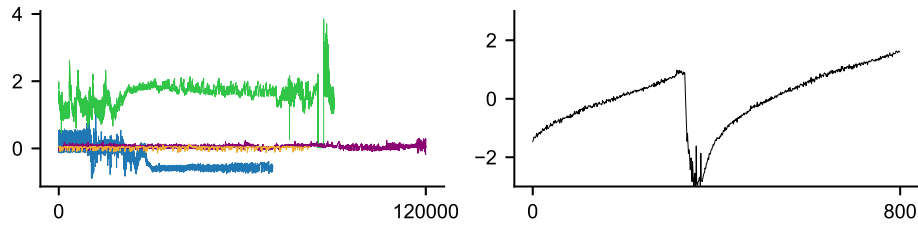
Fig. 2. Four EPG telemetry recordings of Asian citrus psyllid behaviour (*left*). Data set taken from Kamgar et al. [2]. Consensus motif of length 800 (*right*).

FRM-Miner uses the SAX representation [4] to discretise the time series and applies *sequential pattern mining* to them. Then, the discovered discrete patterns are filtered on their overlap with other patterns and patterns with too much overlap are removed. The resulting set of patterns is then mapped back to the continuous time series domain. The occurrences of the patterns are combined to create a *representative* motif. Lastly, the representative motifs are ranked based on the distances between the motif and its occurrences.

Python and C++ implementations, as well as the code and data used for our experiments are available at https://github.com/steenrotsman/frm-miner. Additional code for a subsequent (journal) version will be made available in the same repository.

In order to illustrate the real-world utility of FRM-Miner, we compare its motifs with the consensus motif of a time series database. Fig. 2 shows a collection of four time series that record Asian citrus psyllid behaviour with EPG telemetry. As becomes apparent from the figure, Ostinato finds a well-conserved consensus motif of length 800 in this noisy time series database [2]. With FRM-Miner, we are able to find a representative motif that closely resembles the consensus motif, as well as a previously unknown motif. Both representative motifs are shown in Fig. 3.
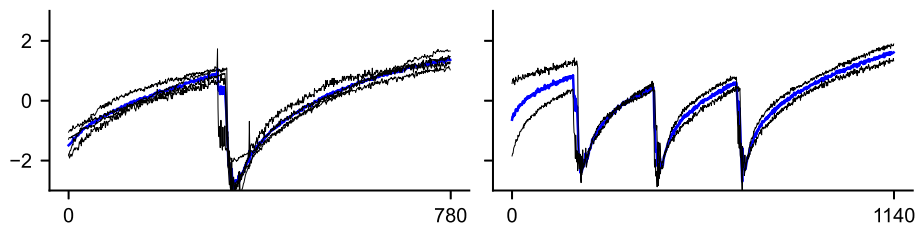
Fig. 3. Representative motif and motif occurrences for insect telemetry. Parameter settings are *minsup* = 1 (left), *minsup* = 0.5 (right), *seglen* = 50, α = 4. The left representative motif corresponds to the motif found by Ostinato, the right motif was previously unknown.

References

1. Feremans, L., Cule, B., Goethals, B.: Petsc: pattern-based embedding for time series classification. Data Mining and Knowledge Discovery **36**(3), 1015–1061 (2022)
2. Kamgar, K., Gharghabi, S., Keogh, E.: Matrix profile xv: Exploiting time series consensus motifs to find structure in time series sets. In: 2019 IEEE International Conference on Data Mining (ICDM). pp. 1156–1161. IEEE (2019)
3. Latecki, L.J., Lakamper, R., Eckhardt, T.: Shape descriptors for non-rigid shapes with a single closed contour. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662). vol. 1, pp. 424–429. IEEE (2000)
4. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing sax: a novel symbolic representation of time series. Data Mining and knowledge discovery **15**, 107–144 (2007)
5. Linardi, M., Zhu, Y., Palpanas, T., Keogh, E.: Matrix profile goes mad: variable-length motif and discord discovery in data series. Data Mining and Knowledge Discovery **34**(4), 1022–1071 (2020)
6. Rotman, S.J., Cule, B., Feremans, L.: Efficiently mining frequent representative motifs in large collections of time series. In: 2023 IEEE International Conference on Big Data (BigData). pp. 66–75 (2023)
7. Torkamani, S., Lohweg, V.: Survey on time series motif discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **7**(2), e1199 (2017)
8. Yeh, C.C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Silva, D.F., Mueen, A., Keogh, E.: Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In: 2016 IEEE 16th international conference on data mining (ICDM). pp. 1317–1322. Ieee (2016)
9. Zhang, M., Wang, P., Wang, W.: Efficient consensus motif discovery of all lengths in multiple time series. In: International Conference on Database Systems for Advanced Applications. pp. 540–555. Springer (2022)
10. Zhu, Y., Yeh, C.C.M., Zimmerman, Z., Kamgar, K., Keogh, E.: Matrix profile xi: Scrimp++: time series motif discovery at interactive speeds. In: 2018 IEEE International Conference on Data Mining (ICDM). pp. 837–846. IEEE (2018)
11. Zhu, Y., Zimmerman, Z., Senobari, N.S., Yeh, C.C.M., Funning, G., Mueen, A., Brisk, P., Keogh, E.: Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins. In: 2016 IEEE 16th international conference on data mining (ICDM). pp. 739–748. IEEE (2016)