Time Series Motif Discovery

Bachelor's Thesis Exposé

eingereicht von:	Jonas Spenger
Gutachter:	Dr. rer. nat. Patrick Schäfer
Gutachter:	Prof. Dr. Ulf Leser
eingereicht am:	10.09.2017

Contents

1 Introduction

Motif discovery is in loose terms the problem of finding interesting patterns in sequences. It is used to discover short and conserved DNA binding sites in the field of bioinformatics, and has also been applied to other domains such as seismic signals and data center coolers. What makes motif discovery interesting is its applicability to a wide range of domains. The difficulty of motif discovery is that the problem of finding interesting patterns is subjective and domain dependent, thus difficult to model mathematically. Consequently existing time series motif discovery methods can be improved upon for both accuracy as well as speed. A graphic example of a time series motif is displayed in figure 1.

Considering the large volume of research and development of DNA sequence motif discovery tools in bioinformatics, it is surprising that these tools have not been applied to

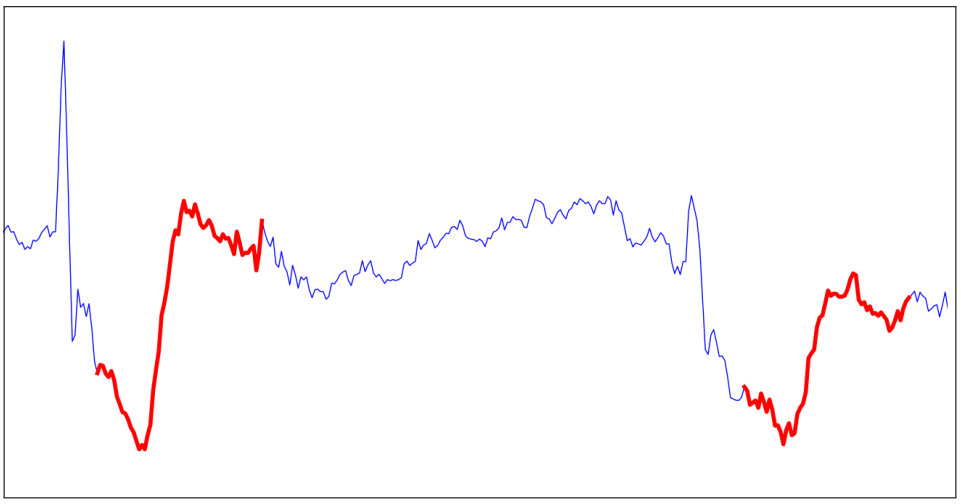

Figure 1: The graph displays an excerpt of a time series consisting of planted ECG measurements (from the UCR Time Series Classification Archive [3]) with added noise. The motif (the two closest subsequences of length 60 measured in z-normalized Euclidean distance) is highlighted in red. Both subsequences occur within a planted ECG measurement and are similar in shape.

time series [1]. These tools are developed to operate on DNA sequences (i.e. sequences over the alphabet 'acgt'). Consequently, they cannot directly be applied to time series (i.e. sequences of real numbers). However, they can be applied to symbolic representations of time series (e.g. a time series discretized and represented as a sequence of characters). Fortunately, there are many preexisting methods for transforming a time series to a symbolic representation. The thesis aims to investigate and evaluate this approach to time series motif discovery.

2 Aims and Objectives

The motivation of the study is to investigate and evaluate how we can apply preexisting DNA sequence motif discovery methods to the time series motif discovery problem. The thesis aims to: 1. Implement a motif discovery pipeline which transforms a time series to a symbolic representation, and thereon applies a DNA motif discovery tool. The goal is to implement a pipeline consisting of ACME [8] (DNA motif discovery method) and SAX [7] (symbolic representation method). 2. Analyze the advantages and disadvantages of

[1] To my knowledge, [4] is the only record in literature of a DNA sequence motif discovery tool applied to time series.

the proposed pipeline. The goal is to compare this pipeline to other time series motif discovery methods on speed, scalability and accuracy.

The ambition is for the proposed pipeline to achieve comparable or better performance than other motif discovery methods (in terms of speed, scalability and accuracy).

3 Background and Related Work

The word *motif* has been defined inconsistently in literature. However, typically motifs are defined as approximately repeated subsequences [2] of a time series [10] [6]. To avoid confusion, the thesis will refer to the following definitions:

Definition 1 *A time series T is a sequence of real numbers $T = (t_1, \ldots, t_n)$.*

Definition 2 *A subsequence S of a time series T is a sequence of real numbers $S = (s_1, \ldots, s_k)$ such that $T = (t_1, \ldots, t_i, s_1, \ldots, s_k, t_{i+k+1}, \ldots t_n)$ for an index i.*

Definition 3 *A motif M of a time series T is a sequence of real numbers $M = (m_1, \ldots, m_k)$ such that there exists at least two subsequences S_1, S_2 of T that are closer than a certain distance d to the motif distance$(M, S_1) < d$ and distance$(M, S_2) < d$ distance (according to some distance function).*

Definition 4 *The top-k motifs are the k highest ranked motifs (according to some rank function).*

Definition 5 *The motif discovery problem is the task of finding the top-k motifs of a time series.*

Motif Discovery: There has been much research on motif discovery. The key difference between motif discovery in the data mining community and motif discovery in the bioinformatics community, is that the former operate on sequences of real numbers (time series) whereas the latter operate on sequences of symbols (DNA sequences).

Recent time series motif discovery research has been on parameter-free and scalable algorithms. Two recently proposed methods are GrammarViz [6], and Matrix Profile [10]. GrammarViz transforms a time series into a symbolic sequence (symbolic representation) using SAX (Symbolic Aggregate approXimation)[7], and thereof infers a

[2] A subsequence (first) is an *approximately repeated subsequence*, if there exists a second subsequence which is within a specified distance threshold to the first subsequence (according to some distance function).

context-free grammar. The set of motifs are generated by the grammar rules. Matrix Profile [10] is a scalable time series motif discovery method with only one parameter. The Matrix Profile [3] is a vector which holds the distance (z-normalized Euclidean distance) of each subsequence to its nearest neighbor. The motifs are found at the lowest value entries of the Matrix Profile vector. Both of these methods can be considered the state of the art in time series motif discovery, and will be used as reference points for the evaluation.

There are many DNA sequence motif discovery methods, such as random projection [2], the MEME Suite [1] and ACME [8]. The thesis will mainly focus on a method named ACME [8]. ACME is a scalable motif discovery method and uses the suffix tree data structure for motif discovery It can find the longest and most frequent approximate matches [4] according to the user specified parameters: frequency, distance, minimum motif length, maximum motif length, etc. ACME is interesting for two reasons: 1. ACME searches for variable length motif candidates, i.e. the search is not restricted to a fixed length. 2. The search for motif candidates is not restricted to subsequences of the input sequence, i.e. better motif candidates can be found as the search space is larger.

For both Matrix Profile and GrammarViz, the search for motif candidates is restricted to the subsequences of the input sequence. However, this is not the case for ACME, as it's search space spans all possible combinations of symbols from the alphabet. Thus, ACME can potentially find motif candidates that are hidden to Matrix Profile and GrammarViz (as the motif candidates are not subsequences of the input sequence). This one of the reasons why we investigate the use of ACME for time series motif discovery in the thesis. In order to apply ACME to time series, a time series must first be transformed into a symbolic representation which will be discussed in the following section.

Symbolic Representation: A popular method for transforming time series into a symbolic representation is Symbolic Aggregate Approximation (SAX) [7]. When applying SAX, the user specifies the alphabet size (i.e. the set of symbols for discretization), and the approximation window (that is the subsequence length that will be converted to a symbol). As the alphabet size increases and the approximation window decreases, the representation converges to the original time series. The runtime of finding approximate motifs using ACME gets considerably slower with increasing alphabet size as the motif discovery search space depends on the alphabet size. The trade-off between accuracy

[3] The *matrix* in Matrix Profile refers to the full distance matrix of all subsequence pairs.

[4] A subsequence (first) is a frequent approximate match, if more than a user specified number of other subsequences of the time series are within a user specified hamming distance to the first subsequence.

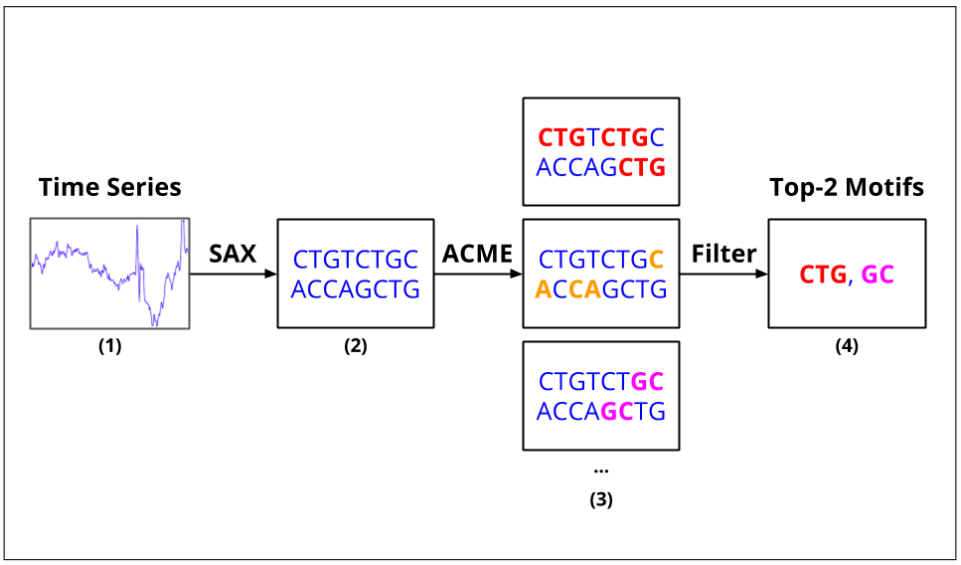

Figure 2: Illustration of the proposed motif discovery pipeline. The intermediate data: (1) Time series; (2) Symbolic representation; (3) Set of motif candidates (in accented colors); (4) Top-2 motifs. The pipeline stages: $(1 \rightarrow 2)$ A time series is transformed using SAX to a symbolic representation; $(2 \rightarrow 3)$ ACME is used to find the set of motifs of the symbolic representation; $(3 \rightarrow 4)$ The set of motifs is ranked and filtered resulting in the top-2 motifs *CTG* and *GC*.

and runtime should be taken into consideration when deciding the alphabet size.

Filtering and Ranking: Filtering and Ranking is usually the last step of the motif discovery algorithm. It is important, as applications such as ACME potentially return a large number of motif candidates. For example, GrammarViz ranks motifs according to length and frequency [6].

4 Method and Results

The proposed motif discovery pipeline will consist of three blocks: SAX, ACME, and Filtering & Ranking. In the first stage, SAX will be used to transform the time series into a symbolic representation. In the second stage, ACME will be used to discover the set of motifs of that symbolic representation. In the last stage, the set of motifs will be ranked and filtered to yield the final top-k motifs. Ranking the motifs will be done according to their length and frequency, such that longer and more frequent motifs have higher rank. Figure 2 is a graphic display of the proposed motif discovery pipeline.

The proposed pipeline will be evaluated and compared to two other well established motif discovery methods, namely Matrix Profile [10] and GrammarViz [6]. The evaluation will be divided into three sections: 1. A case study will be conducted. The retrieved motifs will be visually inspected to see whether there are similarities or differences between the three methods. The subjectivity of motif discovery poses difficulty for the interpretation of the case study. 2. The three methods will be compared on speed and scalability. The runtime for each method on generated random walk time series of varying lengths will be assessed. The dependence on both input and parameters should be taken into consideration when measuring the runtime. 3. The three methods will be compared on accuracy of motif discovery. There is no gold standard for time series motif discovery. However, methods for measuring the motif discovery accuracy have been discussed in [4] [5] [9] and can be adapted for the use in the thesis.

The data used for all the experiments will be the UCR Time Series Classification Archive [3] as well as randomly generated time series. The software needed will be downloaded from the links provided in the respective references.

References

[1] Timothy L Bailey et al. "MEME SUITE: tools for motif discovery and searching". In: *Nucleic acids research* 37.suppl_2 (2009), W202–W208.

[2] Jeremy Buhler and Martin Tompa. "Finding motifs using random projections". In: *Journal of computational biology* 9.2 (2002), pp. 225–242.

[3] Yanping Chen et al. *The UCR Time Series Classification Archive*. `www.cs.ucr.edu/~eamonn/time_series_data/`. July 2015.

[4] Avrilia Floratou, Sandeep Tata, and Jignesh M Patel. "Efficient and accurate discovery of patterns in sequence data sets". In: *IEEE Transactions on Knowledge and Data Engineering* 23.8 (2011), pp. 1154–1168.

[5] Yifeng Gao, Jessica Lin, and Huzefa Rangwala. "Iterative Grammar-Based Framework for Discovering Variable-Length Time Series Motifs". In: *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE. 2016, pp. 7–12.

[6] Yuan Li and Jessica Lin. "Approximate variable-length time series motif discovery using grammar inference". In: *Proceedings of the Tenth International Workshop on Multimedia Data Mining*. ACM. 2010, p. 10.

[7] Jessica Lin et al. “Experiencing SAX: a novel symbolic representation of time series”. In: *Data Mining and knowledge discovery* 15.2 (2007), p. 107.

[8] Majed Sahli, Essam Mansour, and Panos Kalnis. “Parallel motif extraction from very long sequences”. In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM. 2013, pp. 549–558.

[9] Martin Tompa et al. “Assessing computational tools for the discovery of transcription factor binding sites”. In: *Nature biotechnology* 23.1 (2005), p. 137.

[10] Yan Zhu et al. “Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins”. In: *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE. 2016, pp. 739–748.