Motiflets - Simple and Accurate Detection of Motifs in Time Series (VLDB'23)

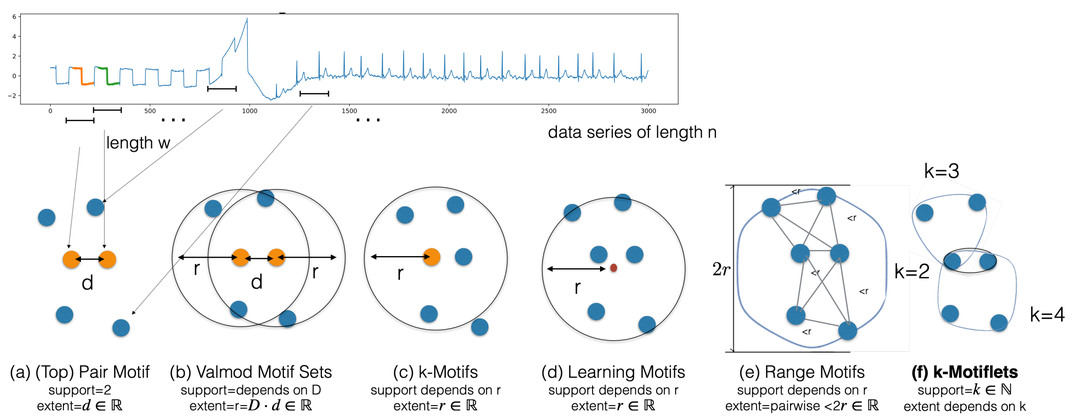
Patrick Schäfer
Humboldt-Universität zu Berlin
Berlin, Germany
patrick.schaefer@hu-berlin.de

Ulf Leser
Humboldt-Universität zu Berlin
Berlin, Germany
leser@informatik.hu-berlin.de

Source Codes

Motif Discovery

- Motifs are **frequently repeated patterns** in time series
- Many definitions exist for the discovery of motifs
- Each one can be mapped to a **geometric shape**

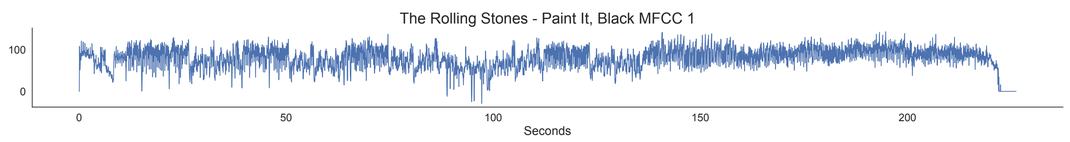


Motif Set Definitions

All definitions have at least **two input parameters**:

- **w: length of motif**
- **r: radius (distance)**

Example: Rolling Stones - Paint it Black

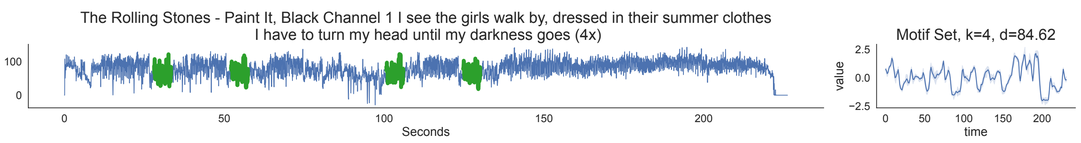


- Motifs correspond to similar rhythmical sections
- This song has a repeating B (8x) and C (10x) part
- We would like to find **all occurrences** of this section (up to 10 repeats)

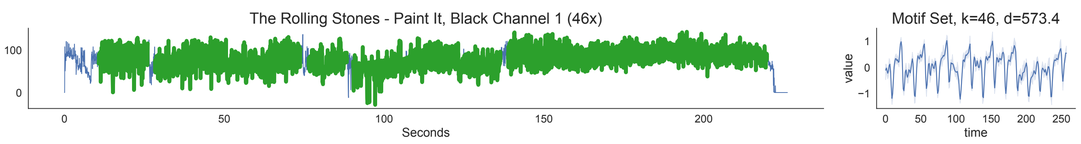
```
I see the girls walk by  (1)
Dressed in their summer clothes
I have to turn my head   (2)
Until my darkness goes
...
```

The Problem

- How to set r and w to find the largest motifs?
- If set **too small**, we miss occurrences



- If set **too large**, we find everything as a motif



- **When are we confident, we found all occurrences?**

Solution: Motiflets

- k-Motiflets is a novel motif discovery algorithm

- It finds k-frequent motifs, and has **two input parameters**:
  - k: size of desired motif set
  - w: length of subsequences

- But no radius/distance as input
- And we can **learn both k and w**

Motif Discovery

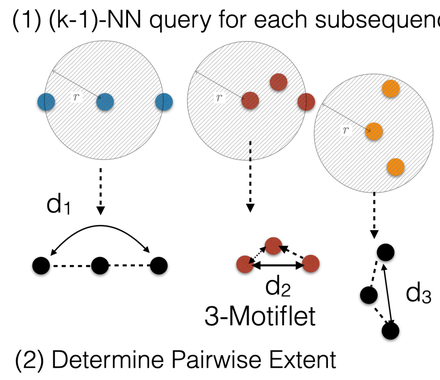
Definitions

The Problem

The Solution

Compute k-Motiflets

1. Perform a k-NN search around each subsequence
2. Compute maximum over pairwise distances ("extent")
3. Minimize over all candidate sets

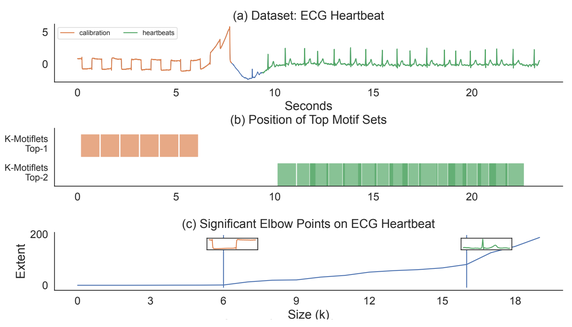


Algorithm 1 Compute Approximate k-Motiflets

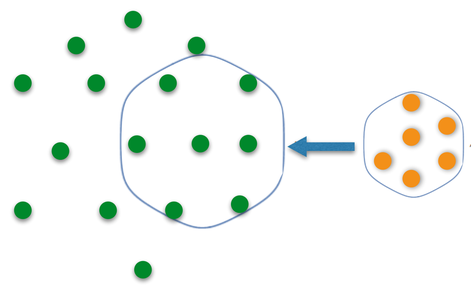
```
1:  procedure GET_K_MOTIFLETS(T, k, l)
2:      D_{i,j} ← CALC_DISTANCE_MATRIX(T, l)
3:      (motiflet, d) ← ({}, inf)                    ▷ bsf of extent
4:      for i ∈ [1, ..., (len(T) − l + 1)] do
5:          idx ← ARGWHERE(D_i < d) ∪ {i}
6:          if len(idx) ≥ k then
7:              candidate ← NON_TRIVIAL_KNN(k, idx)
8:              dist ← PAIRWISE_EXTENT(D, candidate, d)
9:              if dist < d then
10:                 (d, motiflet) ← (dist, k_subset)
11:             end if
12:         end if
13:     end for
14:     return (motiflet, d)
15: end procedure
```

Elbow Plots : Learning k

- We plot the set size k of the k-Motiflet against its extent



Elbow Plot of an ECG trace

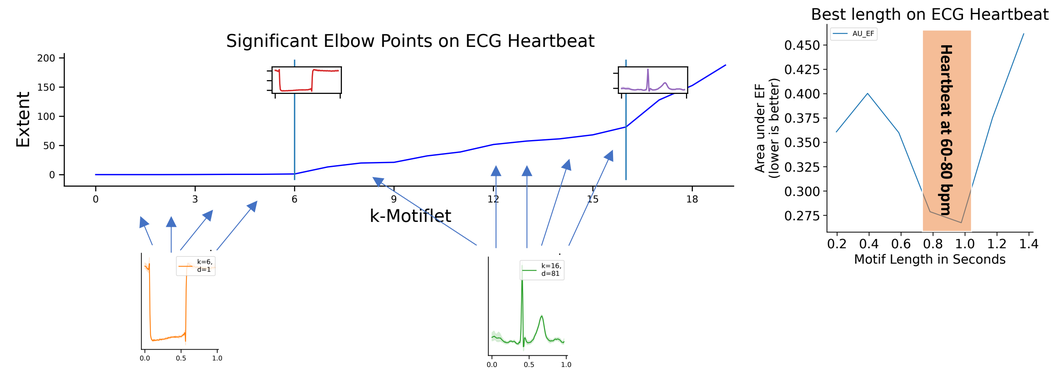


Geometric Interpretation

- Elbows in this plot represent largest meaningful Motif Sets
  - Once increasing k, there is a sharp increase in distances
  - Thus, a dissimilar subsequence is added with large pairwise distance

Area: Learning the window size

- We are interested in flat stretches: these correspond to many repeats
- For each length, we can compute the area under the Elbow Plot
- Minima correspond to long stretches



Evaluation

- We evaluated many real-world use cases against 4 competitors
- Please refer to the paper and supporting website

Algorithm

Learning k

Learning the size