

Multivariate Symbolic Aggregate Approximation for ECG Analysis

Moritz M. Konarski
Supervised by Prof. Taalaibek M. Imanaliev

Applied Mathematics and Informatics Program,
American University of Central Asia

May 3, 2021
Bishkek, Kyrgyz Republic



Outline

1 Introduction

2 SAX and MSAX

3 HOTSAX

4 Preliminary Results

5 Outlook

What is an ECG?

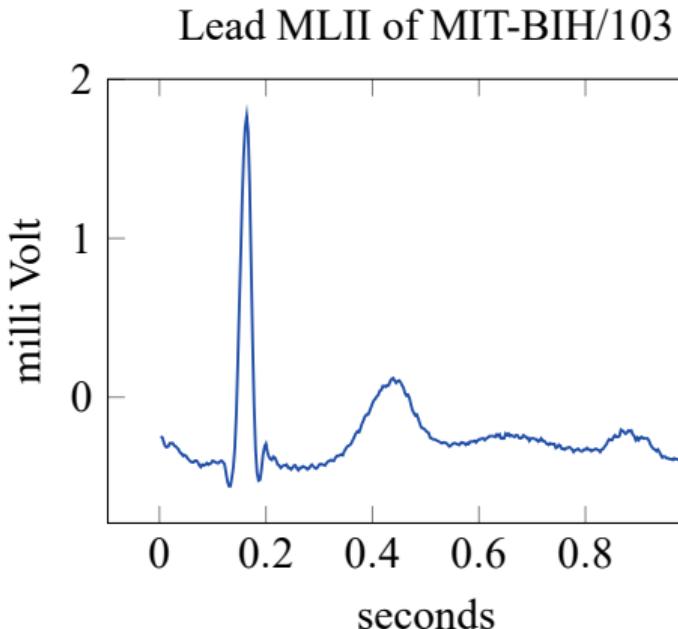


Figure 1: ECG of one heartbeat

- electrocardiogram (ECG or EKG) records the heart's electrical activity
- contains up to 12 simultaneous measurements – the leads
- common medical diagnostic tool

Lead MLII of MIT-BIH/103

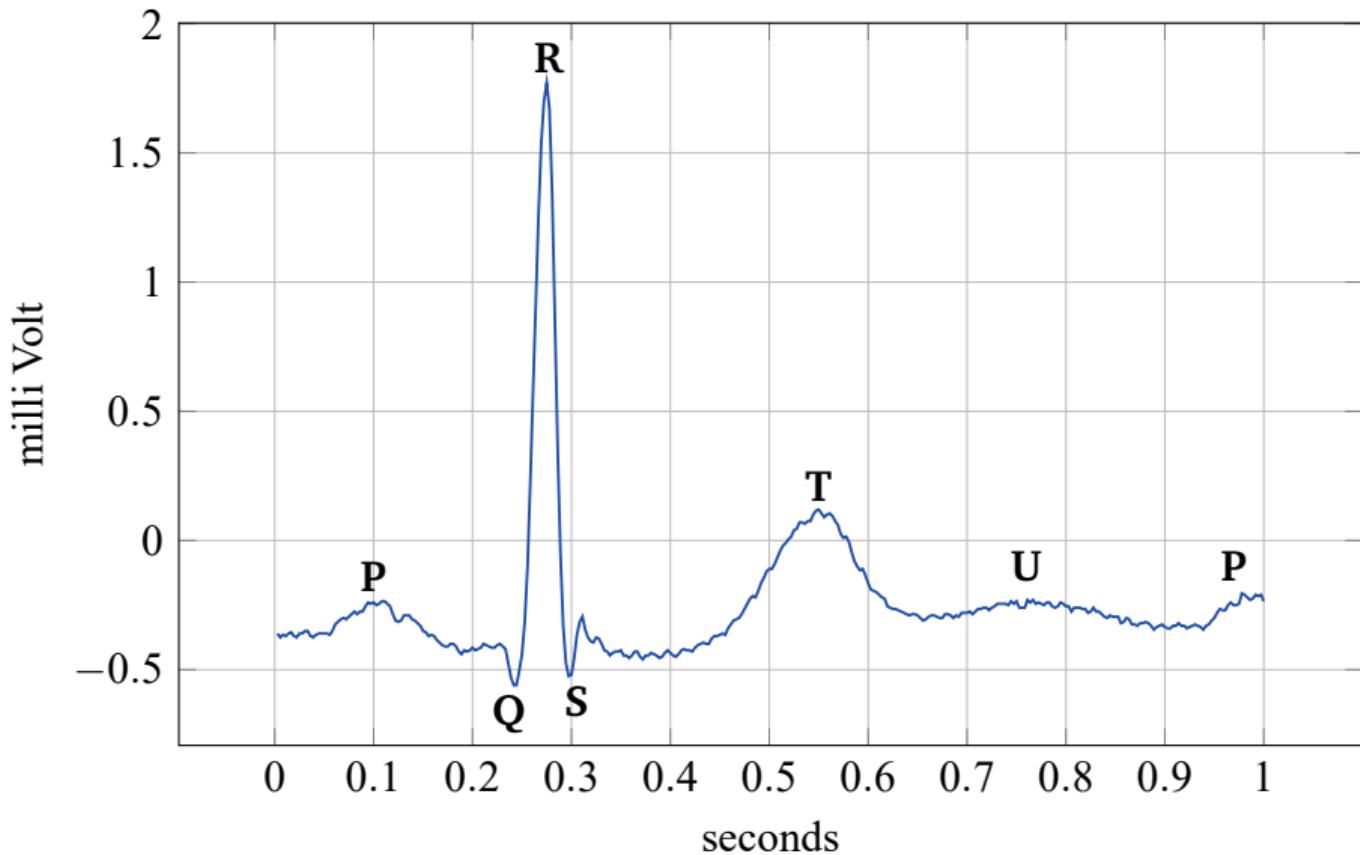


Figure 2: Annotated ECG of one heartbeat

ECGs as Time Series

Definition

A discrete multivariate time series is an ordered sequence at discrete points in time that has n values at each of these points. If $n = 1$, the series is univariate and if $n > 1$, it is multivariate.

- digital ECGs are discrete multivariate time series:
 - have > 1 value at each point
 - recorded at discrete, evenly spaced time points
- time series analysis methods can be applied to ECGs

ECG Analysis

- standard method: manual analysis by cardiologist
- recently: automated (computer-assisted) ECG analysis
- multiple stages:(1) signal acquisition, filtering; (2) data transformation, processing; (3) waveform recognition, feature extraction; (4) classification
- common methods: FFT, DWT, ANN, kNN,...
- relatively new methods are SAX, MSAX, and HOT SAX

SAX, MSAX, and HOT SAX

- Symbolic Aggregate Approximation (SAX) creates a simplified, symbolic representation
 - Multivariate SAX (MSAX) expands SAX to multivariate time series
 - HOT SAX is a discord discovery algorithm that has been used with SAX
- using HOT SAX with MSAX on ECGs should increase the accuracy of discord detection compared to HOT SAX with SAX

Step 1: Z-Normalization

Assumption

The time series values are normally distributed.

SAX

- normalize univariate time series
- uses scalar mean and variance

MSAX

- normalize multivariate time series
- uses vector mean and covariance matrix

⇒ time series have mean 0 and standard deviation 1

Step 2: Dimensionality Reduction

Method

Piecewise Aggregate Approximation (PAA) takes T values and finds the averages of w segments ($w < T$), reducing the complexity.

SAX

- apply PAA to time series

MSAX

- apply PAA to each of the time series individually

⇒ time series has been simplified, consisting of fewer elements

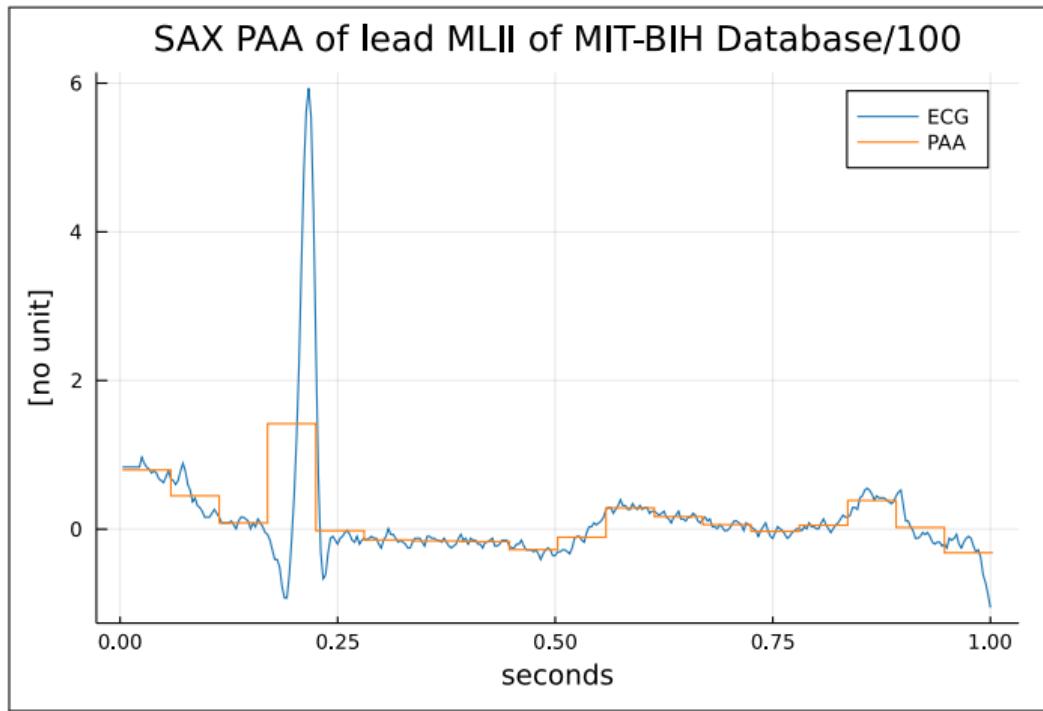


Figure 3: ECG with PAA (MITBIH/100, $w = 18$, $T = 360$)

Step 3: Discretization

Method

Create breakpoints splitting a normal curve into N segments; each segment has equal probability. Then assign a letter to each segment; a to the lowest, b to the next... Result is called a *word*.

SAX

- discretize the time series
- results in one word

MSAX

- discretize each time series individually
- results in one word, multiple letters per segment

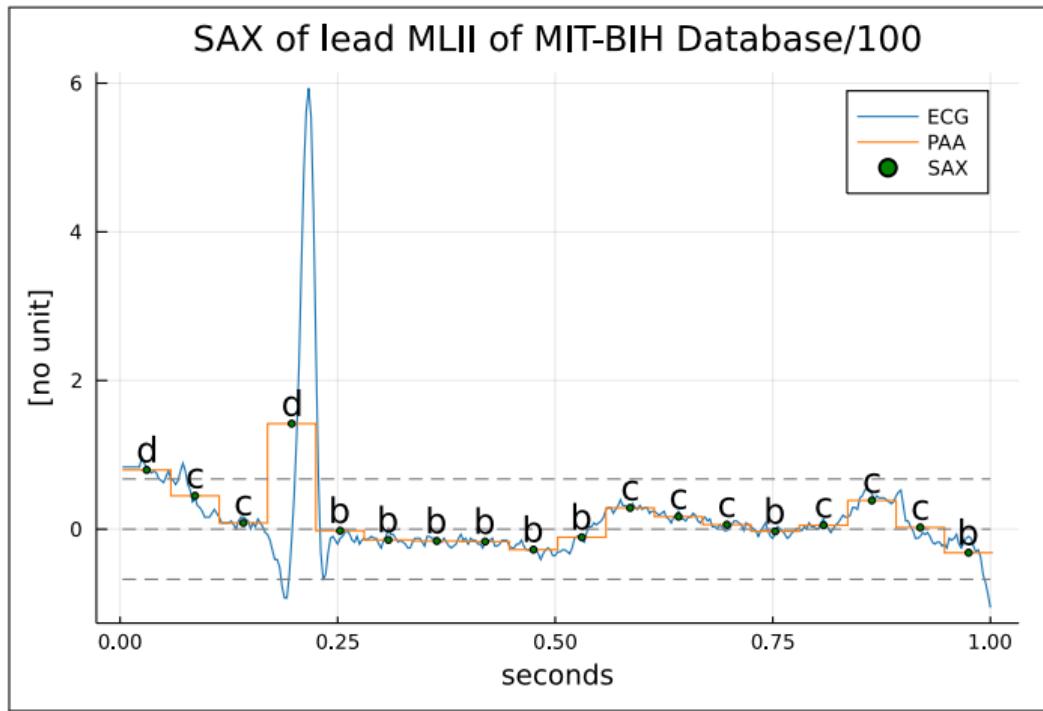


Figure 4: SAX (MITBIH/100, $w = 18$, $T = 360$, alphabet size 4)

Step 4: Distance Measure

Method

To compare two same-length SAX words, a distance measure is needed. Distance is defined for letters: 0 for neighbors; absolute difference of breakpoints otherwise.

SAX

$$\sqrt{\frac{T}{w}} \sqrt{\sum_{i=1}^w (\text{dist}(\hat{q}[i], \hat{c}[i]))^2}$$

MSAX

$$\sqrt{\frac{T}{w}} \sqrt{\sum_{i=1}^w \left(\sum_{j=1}^n (\text{dist}(\hat{q}_j[i], \hat{c}_j[i]))^2 \right)}$$

Time Series Discords

Definition

A time series discord is the subsequence of a time series that is most different from all other segments.

k time series discords are the k most different subsequences.

- discords represent anomalies in an ECG
- can be found by comparing all subsequences to all other subsequences
- for long time series this is not feasible

HOTSAX

- Heuristically Ordered Time series using Symbolic Aggregate Approximation (HOTSAX) is better:
 - discords are rare, start with rarest segment
 - similar segments have similar distances, consider together
- *Hypothesis:* apply HOTSAX and MSAX to ECGs to discover more discords than with HOTSAX and SAX

Preliminary Results

- thus far tested: different alphabet sizes, PAA segment counts, subsequence lengths
- focus mostly on MSAX vs SAX: using MSAX increases the recall rate; while using both SAX curves trumps MSAX
- explain how exactly the checking and stuff works
- used the MIT-BIH ECG database
- database has 48 recordings, every heart beat has been annotated by experts
- use HOTSAX with SAX, MSAX to find discords
- use the annotations to check if the discovered discord is a normal heart

Outlook

- do even more tests with more different set of parameters
- perform more statistical analysis
- explore the influence of parameters
- try this on a 12-lead database to see what happens

Q & A

Thank you!

References I

- [1] G. B. Moody and R. G. Mark, *MIT-BIH Arrhythmia Database*, 1992. doi: [10.13026/C2F305](https://doi.org/10.13026/C2F305).
- [2] M. Anacleto, S. Vinga, and A. M. Carvalho, “MSAX: Multivariate Symbolic Aggregate Approximation for Time Series Classification,” en, in *Computational Intelligence Methods for Bioinformatics and Biostatistics*, P. Cazzaniga, D. Besozzi, I. Merelli, and L. Manzoni, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 90–97, ISBN: 978-3-030-63061-4. doi: [10.1007/978-3-030-63061-4_9](https://doi.org/10.1007/978-3-030-63061-4_9).
- [3] Kligfield Paul *et al.*, “Recommendations for the Standardization and Interpretation of the Electrocardiogram,” *Circulation*, vol. 115, no. 10, pp. 1306–1324, Mar. 2007. doi: [10.1161/CIRCULATIONAHA.106.180200](https://doi.org/10.1161/CIRCULATIONAHA.106.180200).
- [4] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A symbolic representation of time series, with implications for streaming algorithms,” en, in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '03*, San Diego, California: ACM Press, 2003, pp. 2–11. doi: [10.1145/882082.882086](https://doi.org/10.1145/882082.882086).

References II

- [5] C. Zhang *et al.*, “Anomaly detection in ECG based on trend symbolic aggregate approximation,” en, *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2154–2167, 2019, ISSN: 1547-1063. doi: [10.3934/mbe.2019105](https://doi.org/10.3934/mbe.2019105).
- [6] E. Keogh, J. Lin, and A. Fu, “HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence,” en, in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, Houston, TX, USA: IEEE, 2005, pp. 226–233, ISBN: 978-0-7695-2278-4. doi: [10.1109/ICDM.2005.79](https://doi.org/10.1109/ICDM.2005.79).