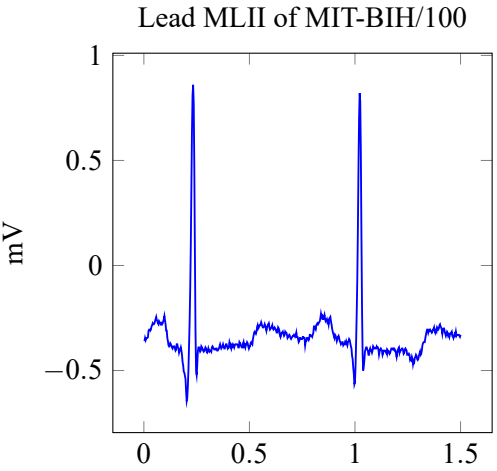
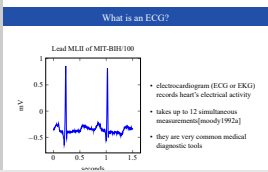
What is an ECG?



- electrocardiogram (ECG or EKG) records heart's electrical activity
- takes up to 12 simultaneous measurements[moody1992a]
- they are very common medical diagnostic tools



- muscle contractions caused by electric pulses
- electric pulse can be measured on the skin
- the measuring things are called electrodes
- electrodes form leads (need 2 to measure anything)
- they have specific positions and names
- 12 leads is the modern standard
- most types of heart disease can be detected
- diagnosis and analysis is performed by trained cardiologists
- datasets available online; contain 2 or more leads (the most significant ones)

ECGs as Time Series

Definition

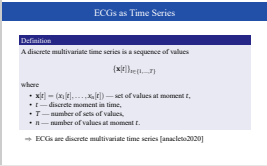
A discrete multivariate time series is a sequence of values

$$\{\mathbf{x}[t]\}_{t\in\{1,\ldots,T\}}$$

where

- $\mathbf{x}[t] = (x_1[t], \ldots, x_n[t])$ — set of values at moment t,
- t — discrete moment in time,
- T — number of sets of values,
- n — number of values at moment t.

$\Rightarrow$ ECGs are discrete multivariate time series [anacleto2020]



- multivariate: measure more than 1 lead per time point
- discrete: set sample frequency in the machines
- discrete: because measured at discrete moments in time
- time series: they are data measured at equal time intervals
- n measurements per point in time (i.e. leads)
- $n = 1$ is univariate, $n > 1$ is multivariate

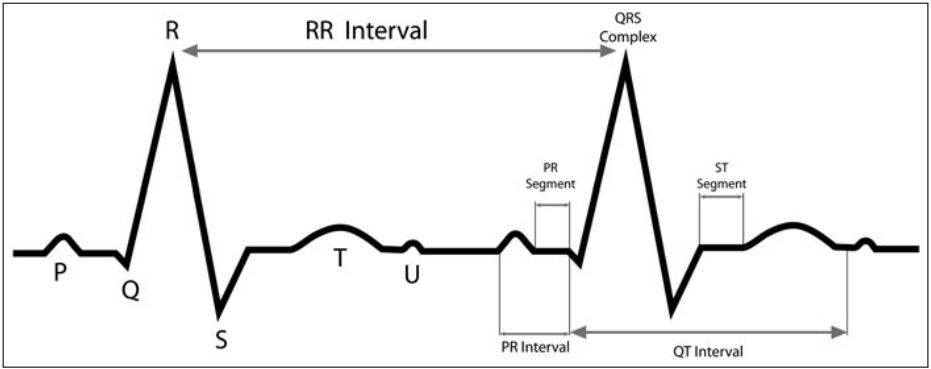


Figure 2: An annotated model ECG [wasilewski2012a]

Automated ECG Analysis

- ECGs represent large amounts of data, thorough analysis is required
- 5 stages: (1) signal acquisition, filtering; (2) data transformation, processing; (3) waveform recognition; (4) feature extraction; (5) classification[kligfieldpaul2007]
- some methods include FFT, DWT, ANN, kNN, filters
- balance between accuracy and complexity needed

SAX and MSAX

- Symbolic Aggregate Approximation (SAX) creates a simplified, symbolic representation [lin2003]
- is guaranteed to behave like the original data
- works on univariate time series, has been used on ECGs[zhang2019]
- Multivariate SAX (MSAX)[anacleto2020] expands SAX to multivariate time series

$\rightarrow$ using MSAX on ECGs should increase the accuracy of discord detection compared to SAX

Z-Normalization

- assumption: data is approximately normally distributed
- to analyze time series, they are first normalized so that $\mu = 0$ and $\sigma = 1$

$$x^i[t] = \frac{X^i[t] - \mu}{\sigma} \tag{1}$$

- enables comparison between different time series

PAA

- piecewise aggregate approximation (PAA) reduces dimensionality (through averaging of segments)
- simplifies the time series
- results in $\bar{C} = \bar{c}_1, \ldots, \bar{c}_w$
- getting element i of $\bar{C}$ (time series x has length n)

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} x_j$$

PAA Graph

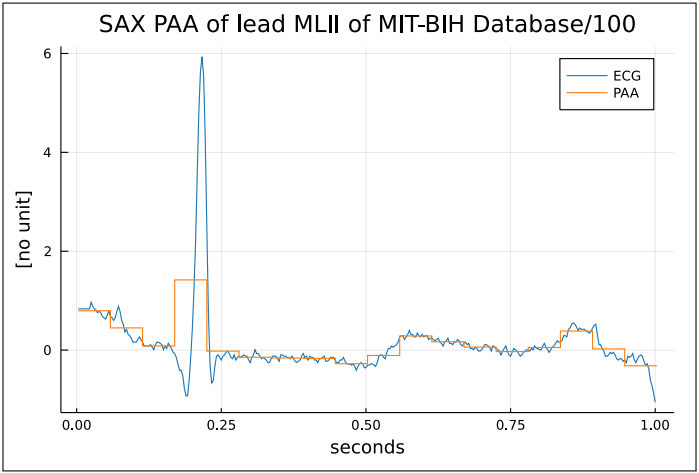


Figure 3: ECG with PAA (MITBIH/100, $w = 18$, $n = 360$)

Discretization

- assign letters to PAA segments
- breakpoints are created that divide a Gaussian curve into equal parts
- number of breakpoints dependent on size of alphabet
- all PAA below lowest breakpoint are a, the ones above it b…

Discretization Graph

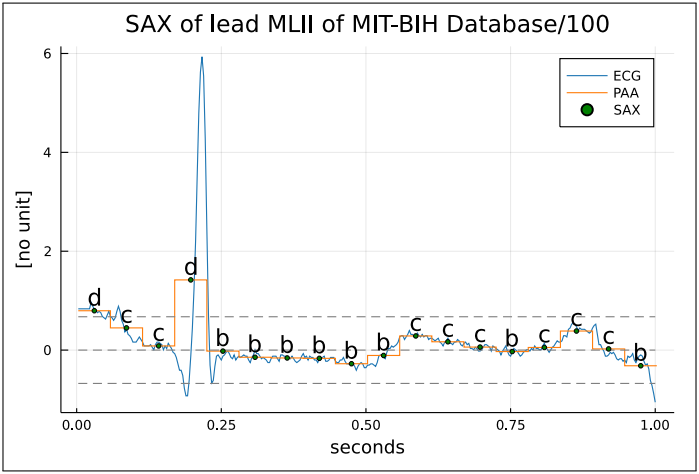


Figure 4: SAX (MITBIH/100, $w = 18$, $n = 360$, alphabet size 3)

Distance Measure

- SAX lower bounds the Euclidean distance, i.e. SAX distances correspond to Euclidean distances
- Euclidean distance between 2 time series Q, C

$$D(Q, C) \equiv \sqrt{\sum_{i=1}^{n} (q_i - c_1)^2}$$

- SAX distance

$$MINDIST\left(\hat{Q}, \hat{C}\right) \equiv \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} \left(dist(\hat{q}_i, \hat{c}_i)\right)^2}$$

- $dist(\hat{q}_i, \hat{c}_i)$ is the difference between the breakpoints of $\hat{q}_i, \hat{c}_i$

Z-Normalization

- perform multivariate normalization
- mean vector μ as vector of the means for each time series
- covariance matrix Σ for variances and covariances between the different time series

$$\mathbf{x}[t] = \Sigma^{-1/2}(\mathbf{X}[t] - \mu)$$

- this uses mean and covariance structure of the multivariate data

PAA and Discretization

- PAA is used here like in SAX, each time series is handled separately
- the discretization process works the same way too
- each time series component is discretized separately
- to differentiate them, one alphabet can for example be uppercase

Discretization Graph

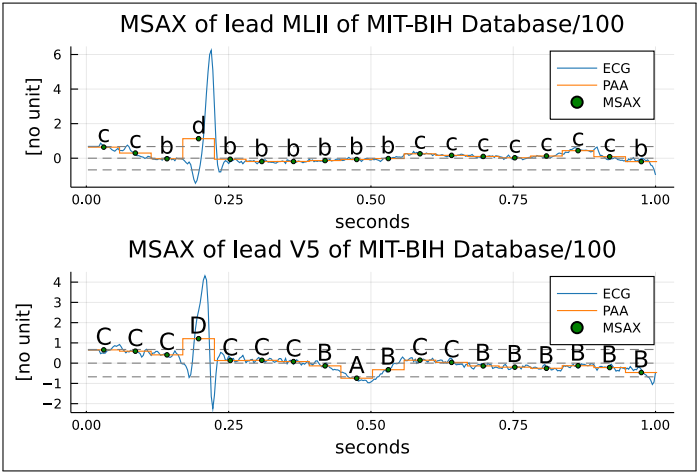


Figure 5: MSAX (MITBIH/100, $w = 18$, $n = 360$, alphabet size 3)

Distance Measure

- this distance measure is based on *MINDIST*
- it is also lower bounding the Euclidean distance
- it adds an extra step of adding the *dist* values for the time series components

$$MINDIST\_MSAX(Q,C) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=0}^{w} \left(\sum_{i=0}^{n} dist(q[i], c[i])^2 \right)}$$