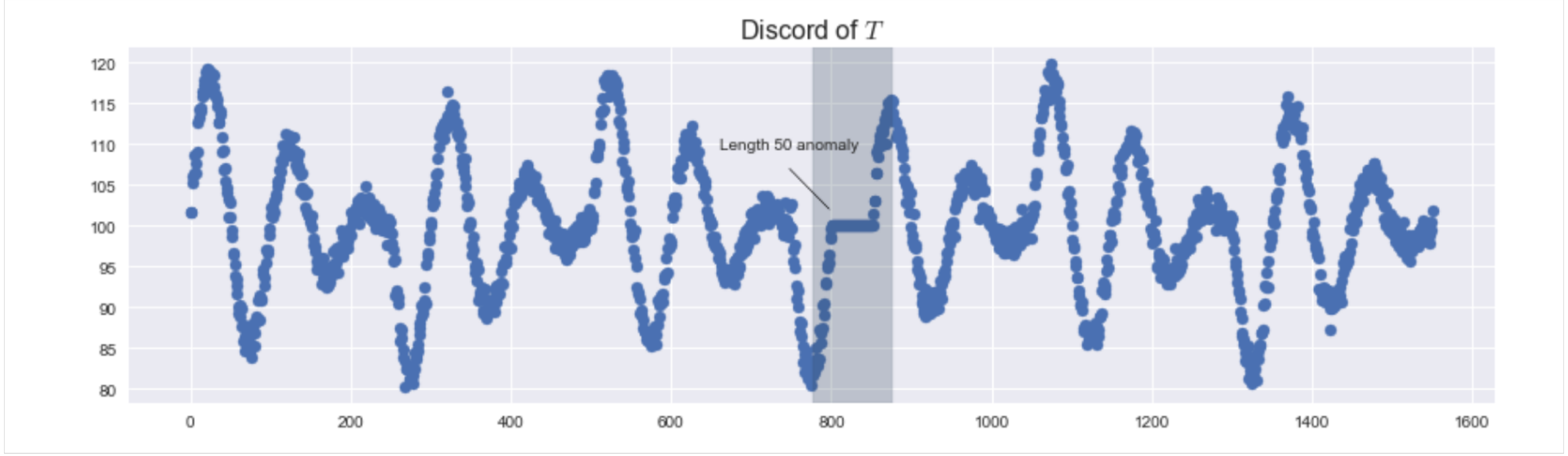
Grammar Compression

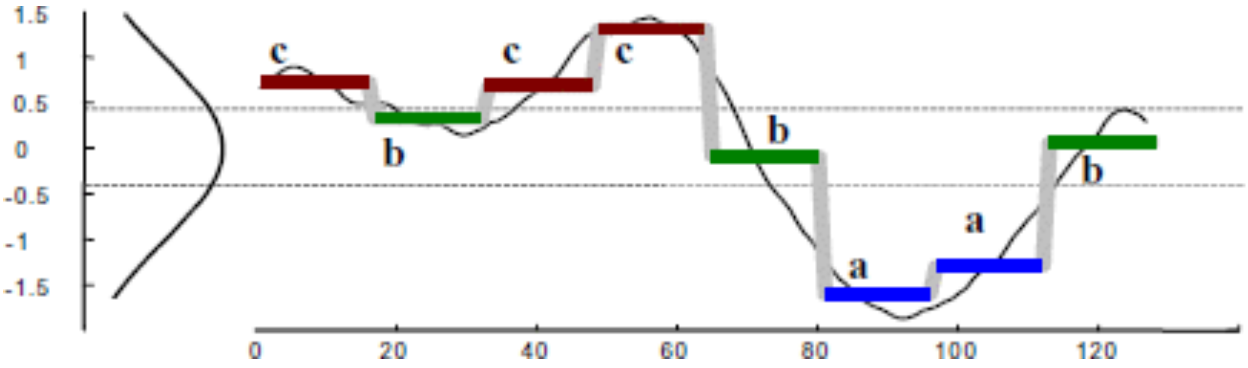
Brute Force Algorithm



- Brute force algorithm:
    - Numerically computes furthest nearest neighbor
    - Not vectorized (in literature comparison; can be vectorized)

Symbolic Aggregate approXimation

SAX



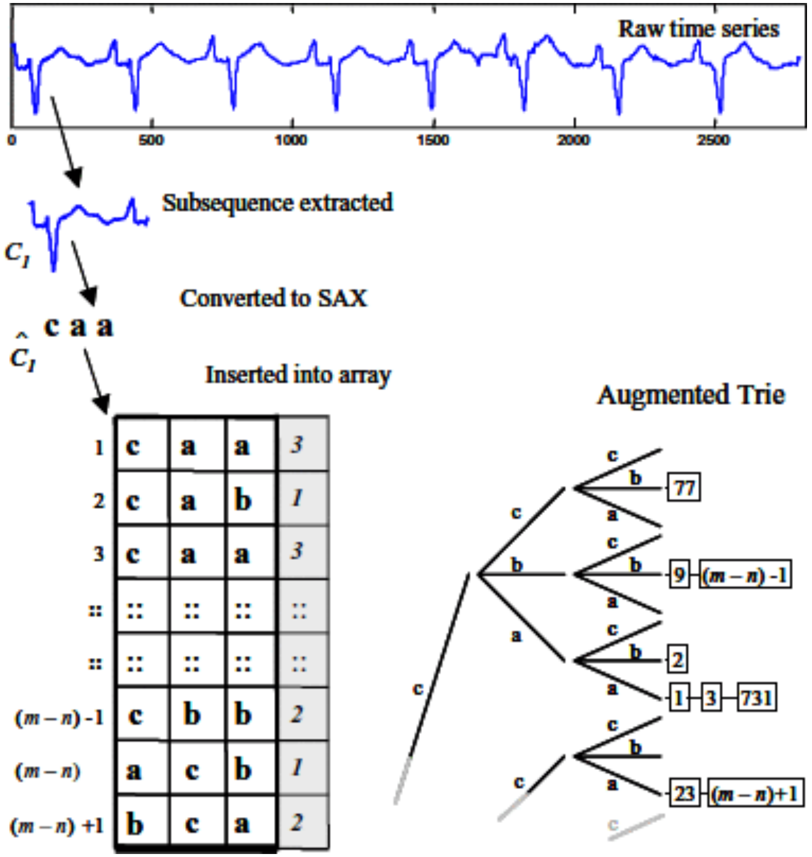
- Converts data subsequences to letters using Piecewise Aggregate Approximation (PAA)
- Letters are assigned based on pre-selected (pseudo-arbitrary) alphabet boundaries
- Series of letters (usually 3-4) are combined into words

Anomaly Detecition with SAX

- Check distance between words (using modified distance function)

Heuristically Ordered Time Series SAX

HOT SAX



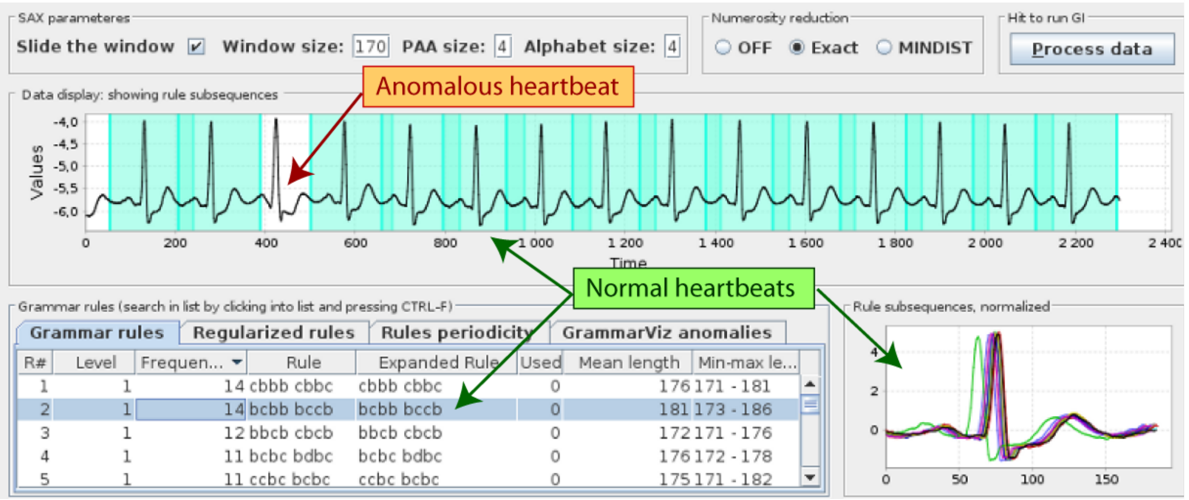
1	c	a	a	3
2	c	a	b	1
3	c	a	a	3
::	::	::	::	::
::	::	::	::	::
(m − n) -1	c	b	b	2
(m − n)	a	c	b	1
(m − n) +1	b	c	a	2

- Improves SAX by favorably looking up ″rare″ words first
- Reported 3,000x speedup on brute force algorithm
- HOT SAX not vectorized
  - doing so may interfere with heuristic time savings?
  - I think it could be partially done (if necessary)…

Anomaly Detection with HOT SAX

- Words are saved in a lookup trie
- Rare words suggest anomalies; then distance is checked between words to confirm

GrammarViz



SAX parameteres

Slide the window ☑ Window size: 170 PAA size: 4 Alphabet size: 4

Numerosity reduction

○ OFF ● Exact ○ MINDIST

Hit to run GI

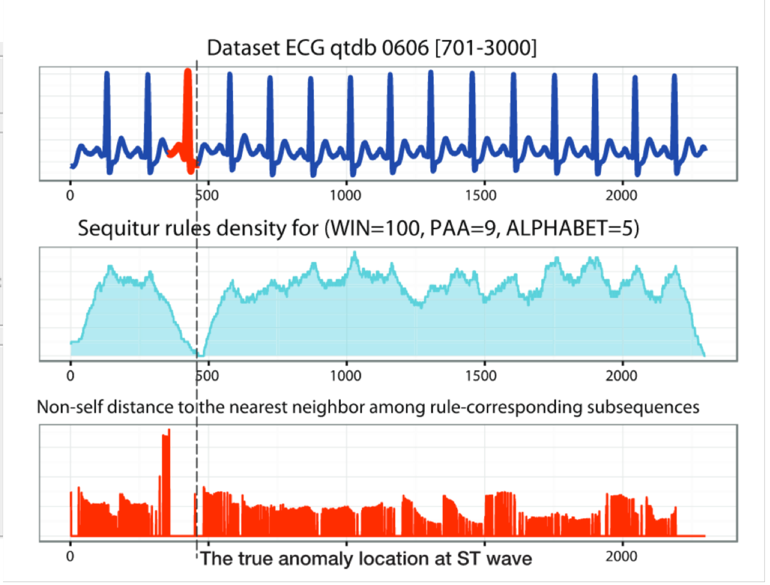
Process data

Data display: showing rule subsequences

Grammar rules (search in list by clicking into list and pressing CTRL-F)

Grammar rules | Regularized rules | Rules periodicity | GrammarViz anomalies

R#	Level	Frequen...	Rule	Expanded Rule	Used	Mean length	Min-max le...
1	1	14	cbbb cbbc	cbbb cbbc	0	176	171 - 181
2	1	14	bcbb bccb	bcbb bccb	0	181	173 - 186
3	1	12	bbcb cbcb	bbcb cbcb	0	172	171 - 176
4	1	11	bcbc bdbc	bcbc bdbc	0	176	172 - 178
5	1	11	ccbc bcbc	ccbc bcbc	0	175	171 - 182



Dataset ECG qtdb 0606 [701-3000]

Sequitur rules density for (WIN=100, PAA=9, ALPHABET=5)

Non-self distance to the nearest neighbor among rule-corresponding subsequences

The true anomaly location at ST wave

- Uses **Rule Density Curve** and **Rare Rule Anomaly (RRA)** as detection algorithms
- **Rule Density Curve:** Number of grammar rules covering each point in time series (does not explicitly calculate distances between subsequences)
- **RRA:** follows procedure of HOT SAX but using grammar to inform the heuristic