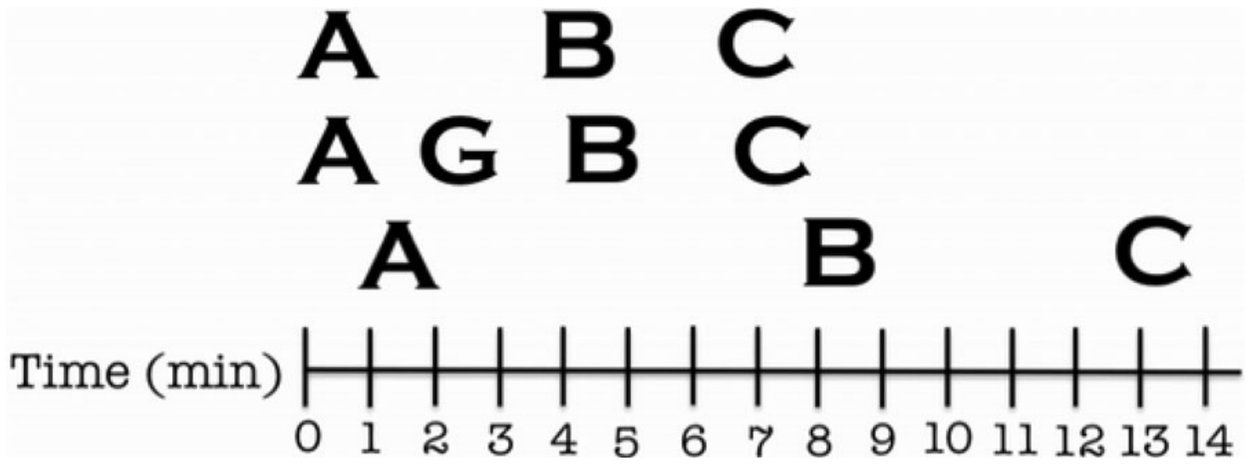
Comparing finite sequences of discrete events with non-uniform time intervals

Alexander Murph, Abby Flynt, and Brian R. King

*A **NEW** SIMILARITY SCORE FOR AN **OLD** PROBLEM*

Sequences of events in time show up in many disparate fields. Clustering and classifying these sequences requires a notion of similarity/distance that can take into account *both* sequence similarity and *when* these events take place.



Comparing sequences of events in time may require more than basic alignment algorithms. Here, when time is considered, the "most alike" sequence changes.

The Algorithm

The *Sequence Alignment with Non-uniform Time Intervals* (SAWNUTI) algorithm extends the well-known Smith-Waterman (for local alignment) and Needleman-Wench (for global alignment) sequence comparison algorithms.

Algorithm 2: Sequence Alignment with Non-Uniform Time Intervals

Input: $\Phi = \phi_1\phi_2 \dots \phi_m$ and $\Theta = \theta_1\theta_2 \dots \theta_n$: sequences over $\mathcal{S}$
$T = t_0\ t_1\ \dots\ t_{m-1}$ and $\mathcal{T} = \tau_0\ \tau_1\ \dots\ \tau_{n-1}$: time intervals for Φ and Θ, respectively
$\mathbb{T}$: Every time interval in the entire data set of interest
α : time interval bias
$sim(\phi, \theta)$: similarity function
$W(x)$: penalty function for gap of length $x \in \mathbb{N}$

Output: scoring matrix H

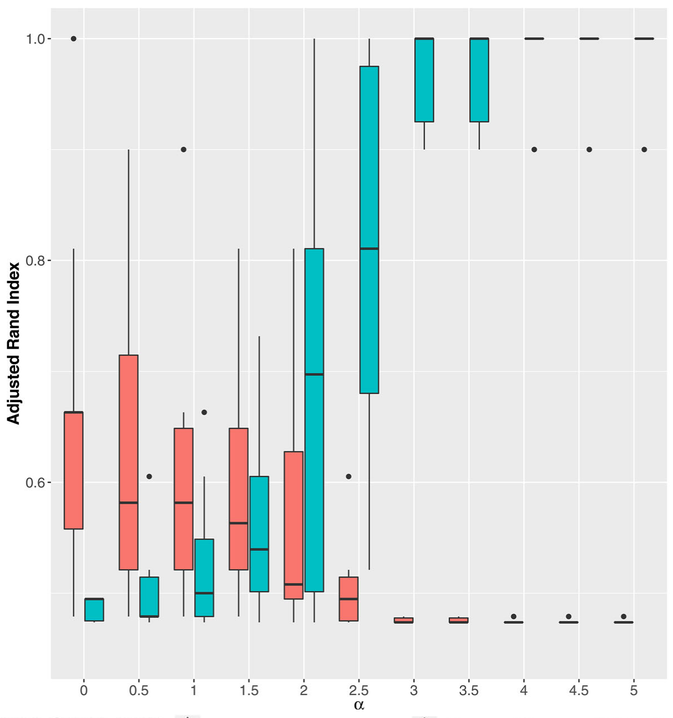
$H \longleftarrow (m+1) \times (n+1)$ matrix;
Initialize first row and column of H;
Transform each sequence into a sequence of time intervals;
Perform the zero-one transformation: $\forall \hat{t} \in \mathbb{T}, \frac{\hat{t} - min(\mathbb{T})}{max(\mathbb{T}) - min(\mathbb{T})}$;
for $i \leftarrow 1...m$ **do**
  for $j \leftarrow 1...n$ **do**
    k = gap in sequence Φ;
    l = gap in sequence Θ;
    $\Phi_{k,l}(i,j) = |(t_i + t_{i-1} + \dots + t_{i-k}) - (\tau_j + \tau_{j-1} + \dots + \tau_{j-l})|$
    if *Global Alignment* **then**

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} - \Phi_{0,0}(i-1, j-1) * \alpha + sim(\phi_i, \theta_j) \\ H_{i-k,j} - \Phi_{k,l}(i-1, j) * \alpha - W(k) \\ H_{i,j-l} - \Phi_{k,l}(i, j-1) * \alpha - W(l) \end{cases}$$

    else

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} - \Phi_{0,0}(i-1, j-1) * \alpha + sim(\phi_i, \theta_j) \\ H_{i-k,j} - \Phi_{k,l}(i-1, j) * \alpha - W(k) \\ H_{i,j-l} - \Phi_{k,l}(i, j-1) * \alpha - W(l) \\ 0 \end{cases}$$

    end
  end
end



ARIs for hierarchical clustering (Ward's linkage) on distinguishing the underlying Markov chain and time distributions. Experiment was replicated 10 times for each value of time interval bias α.

Classification on Simulated Data

i. To simulate our synthetic data ***sequence distributions***, we generated sequences of varying length from two Markov processes (Markov A & Markov B). Each process steps through one of 5 states with starkly different transition probabilities.
ii. To simulate our synthetic data ***time distributions***, we drew a time value for every "gap" available in the sequences from (i). These came from two starkly different uniform distributions (Narrow & Wide).

The above process gives four data subgroups:

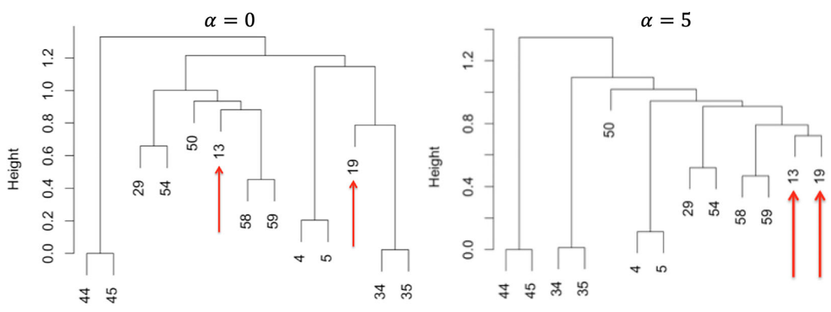
- Markov A, Narrow
- Markov A, Wide
- Markov B, Narrow
- Markov B, Wide

These were generated such that we have an equal number of sequences from each of the four data subgroups.

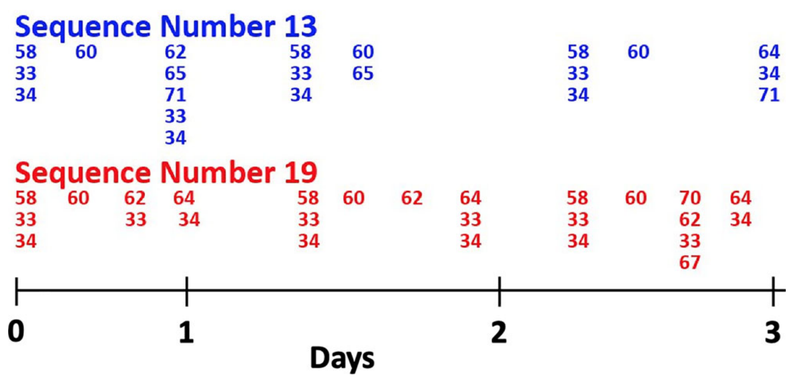
Application using Classification

We obtained a medical dataset of diabetes patients who were logging their condition and their behavior related to managing their diabetes over time. See: UCI ML repository (https://archive.ics.uci.edu/ml/datasets/diabetes)

I. Each sequence in this dataset corresponds to a single patient's record of events over time; each event has a numerical code.
II. Possible events include things like insulin doses, glucose measurements, meals, and exercise.
III. Data includes records that use four "timestamps"; breakfast = 08:00, lunch = 12:00, dinner = 18:00, bedtime = 22:00; as well as more specific times. These were all gathered into a single data set.



Dendrograms for clustering sequences of events in time of diabetes patients. When time is important, patients 13 and 19 are clustered together much sooner.



The events of patients 13 and 19 mapped over the course of 3 days.

Discussion

i. The SAWNUTI algorithm **extends two existing sequence alignment algorithms** to allow for the additional variable of time.
ii. SAWNUTI **performed strongly on simulated data**, much better than competitors (see: ScanMatch alg).
iii. Looking at data over time of diabetes patients, **we can see how much of a difference time makes when comparing two patients**.
iv. **See paper for an additional study** comparing sequences of eye tracking fixations of autistic children on the same stimulus over time.

Acknowledgements

We express our gratitude to Dr. Vanessa Troiani and Dr. Antoinette Dicrisco at Geisinger Autism and Developmental Medicine Institute in Lewisburg, PA, for their interest in our work, for their insight, and for contributing eye tracking data.

Citation:

A. Murph, A. Flynt, B. R. King (2021). Comparing finite sequences of discrete events with non-uniform time intervals, *Sequential Analysis*, 40(3), 291-313.