SAGA: A novel signal alignment method based on genetic algorithm

Hüseyin Kaya [a,*], Şule Gündüz-Öğüdücü [b]

[a] *Istanbul Technical University, Informatics Institute, Istanbul, Turkey*
[b] *Istanbul Technical University, Computer Engineering Department, Istanbul, Turkey*

ARTICLE INFO

ABSTRACT

Signal alignment is one of the most commonly used strategies in analyzing a group of time series in order to learn the variations or common patterns across individual signals. A pair-wise alignment algorithm aligns two signals by warping the time axis of the first signal so that the warped signal is "near" to the second. The majority of alignment algorithms are focused on extracting features like the locations of significant peaks or peak widths, and using those features in aligning the signals instead of raw signal. Although this approach allows fast alignments, it suffers from the risk of missing important features, leading to inaccurate alignments. In this paper, a novel **S**ignal **A**lignment method based on **G**enetic **A**lgorithm (SAGA) is proposed to align raw signals by first modeling the warping function with an ODE model. The parameters of the warping function are then optimized by using a genetic algorithm. The SAGA does not require feature extraction and it preserves the smoothness of the signals. The performance of the proposed method is evaluated on two sets of synthetic and real world datasets and compared to the well-known alignment algorithms. The results show that SAGA is a powerful algorithm that can compete with the others.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Time series is one of the common data types encountered in many applications. It can be obtained by periodically measuring the height of a child [19] or it can be an outcome of a sophisticated device such as a DNA sequencer [24] or a Mass Spectrometer [6]. In either case, a time series is a list of measurements obtained by observing an event over time. The variations or patterns in the data usually carry useful information about the investigated subject. For example, height measurements describe the growth rate of children or the data produced by a DNA sequencer indicate the DNA sequence of a patient under inspection.

Processing a single time series and extracting useful information out of it are well-known research topics and deserve attention on their own. A simple example is to create a 4-nucleotide sequence by analyzing the time series provided by DNA sequencers [9]. However, a more interesting problem is to analyze a group of time series in order to learn the variations or common patterns across the individuals. Extracting that information about the group may provide better understanding on the investigated study. Analyzing a set of growth profiles of male and female children [27], detecting changes in peptide/protein abundances in samples [2], digital tracking of lesions [13] or comparing a group of DNA sequences to a reference [5] are such examples.

Many techniques used in analysis of multiple time series assume that they are ready to make a side by side comparison. For example one may want to cluster the time series into distinct groups by assuming that the location of a peak common in all time series are at the same time point. However shifts in time series is a common phenomena in many application fields. Shifts, which are also called time drifts or retention time difference, can be highly non-linear because of imperfections of measurement device or differences in the examined subjects. Therefore it is often necessary to correct the time drifts between the time series. This involves stretching or compressing time axis of one or more time series so that the resulted series are ready to analyze. This is called an alignment.

There are two basic strategies in alignment of a group of time series. In the first approach, alignment is carried out by using the features extracted from raw data. These features used to align original raw data. Conventional mutation screening fits very well to this description [15,16]. The raw data (signal) generated by DNA sequencers are first processed to obtain a 4-nucleotide DNA sequence (features), then these string sequences are aligned. Another application field is the alignment of chromatographic profiles which stems from Gas/Liquid Chromatograph-Mass spectrometry (MS) platform for the profiling of certain classes of small molecules in biological samples. In this type of data, the locations of prominent peaks are called feature vectors which are generally obtained by finding the points at which the first and second derivatives are zero and positive, respectively. Feature vectors can then be aligned by various procedures including Dynamic Programming (DP) [21] and fuzzy matching approach [29]. A detailed review on signal alignment based on extracted features can be found in [1]. The advantage of this approach shows itself in the name of speed since the transformation reduces data size which allows a faster analysis. However, it also has an important drawback; the risk of missing a feature, resulting in misleading conclusions in later steps. For example, during the generation of DNA sequences, a single false nucleotide assignment may lead to an erroneous diagnosis of the patient.

The second strategy depends on the idea of using raw data throughout the entire comparison study in order to eliminate the risk of incorrect feature extraction of the first approach [7,14,15]. The alignment of raw signals is achieved by "warping" the time axis of one or both of the signals so that the two signals are forced to be "close" to each other in the sense that the Euclidean distance between them is small. Warping is defined as nonlinear stretching and compressing the signal along the time axis.

The clock analogy is very useful to further understand this phenomena. Consider two people observing the same event and recording the outcomes by using the time displayed in their individual clocks, namely A and B, as shown in Fig. 1. If there is no time difference between A and B, then the recorded observations will be identical. However, there will be differences between the signals depending on the amount of time drifts which can be modelled by a strictly monotone increasing function, which is called a "warping function", mapping the time in A to the time in B. If one knows the warping function, then it can be used to align the two signals by applying its inverse only to the second signal. In this way, the time axis of the second signal is corrected to match that of the first. In other words, we can say that only clock B is adjusted to create a synchronization between the clocks. Likewise, if one applies the warping function to the first signal and keeps the second signal fixed, then the time axis of the first signal is corrected to match that of the second. This is equivalent to adjusting only the first clock A to create a synchronization. These two cases are examples of asymmetric alignment since only one of the signals is warped. In symmetric cases both signals are warped to create an alignment. This is equivalent to making adjustments in both clocks. The symmetric case is further explained in Section 3.

Aligning a pair of signals, as in the above example, is called a "pairwise alignment". Aligning more than two signals is called a "multiple alignment" and it can be achieved by doing a series of pairwise alignments. One of the signals can be chosen as a "reference", then the other signals are pairwise aligned to the reference one at a time. The average of the signals can be chosen as a reference. It is also possible to create a representative signal by assuming that each time series is generated as a noisy transformation of a single signal called a "latent trace" which can be obtained by Continuous Profile Model (CPM) [12]. CPM has difficulties in aligning signals longer than 1000 points and not meaningful for just a pair of signals. Therefore it is not used for comparison purposes in the experimental part of this study. From now on, an alignment represents pairwise alignment unless otherwise stated.

The stated optimization problem to the signal alignment has been a major interest in the community for nearly half a century. The vast majority of the methods developed in this era are built on Dynamic Programming (DP) [4] that lends itself to the creation of cost matrices which becomes a liability in the alignment of long signals. It also generates non-smooth warping functions resulting in sudden jumps in the alignment. Many attempts have been made to solve these problems, but none of them achieved both objectives.

(a) Two different observations of an event   (b) Asymmetric alignment   (c) Symmetric alignment
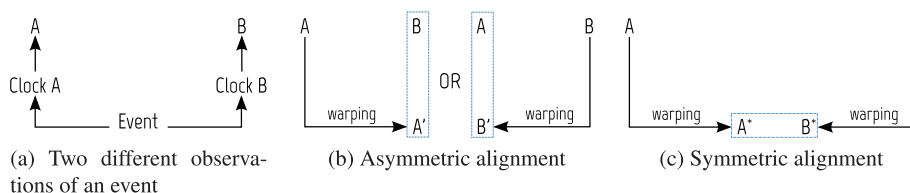
Fig. 1. (a) Same event is measured by two independent observers with clocks A and B, respectively. Resultant time series are A and B. (b) In asymmetric alignment, warping is only applied to one of the signals. (c) In symmetric case, both time series are warped.