Aligning gene expression time series with Invariance to Uniform Scaling with Multiple Scaling Factors

Coleman Yu & Tatsuya Akutsu

Bioinformatics Center, Institute for Chemical Research, Kyoto University

{cyu, takutsu}@kuicr.kyoto-u.ac.jp

Abstract

Biological processes can be analyzed by using the time series manner. For example, biological processes that develop over time can be studied by collecting RNA expression data at selected time points and the distinct cycles can thus be identified [1]. In other to achieve it, we need to have a similarity measure subroutine that can return the similarity between two series that matches our intuition. In other words, this subroutine allows us to find a data point of one series that maps to the data point on other series in a good way. However, biological data is intrinsically noisy. Besides, even for a the same process in different conditions, the instances of it may unfold at different rates. Hence, we cannot simply using the Euclidean distance to measure the similarity between two series: Simply pair off the points taken at the same time rate. The sequences must be warped in a nonlinear fashion to match each other. This can be achieved by a method called "**D**ynamic **T**ime **W**arping (DTW)" [1]. Recent studies shows that DTW is the best measure in most domains.

However, if the two series are in different scales in the y-axis, even DTW cannot returns the convincing result. The scaling invariance can be achieved by using a method called "Uniform Scaling (US)". It is natural to use the combination of DTW and US (USDTW) to handle both the local invariance and the global invariance [2]. In this poster, we will argue that even USDTW cannot handle some cases. The one drawback of the original uniform scaling definition is that it only employs one scaling factor, However, a process may consists of different sub-processes. Each of them has it own unfold rate. Multiple scaling factors instead of a single scaling factor are required to handle this scenario. We will introduce an improved version of USDTW that leverages of multiple scaling factors to return a better result.

The following sections present the basic knowledge to understand this poster.

What is Time Series?

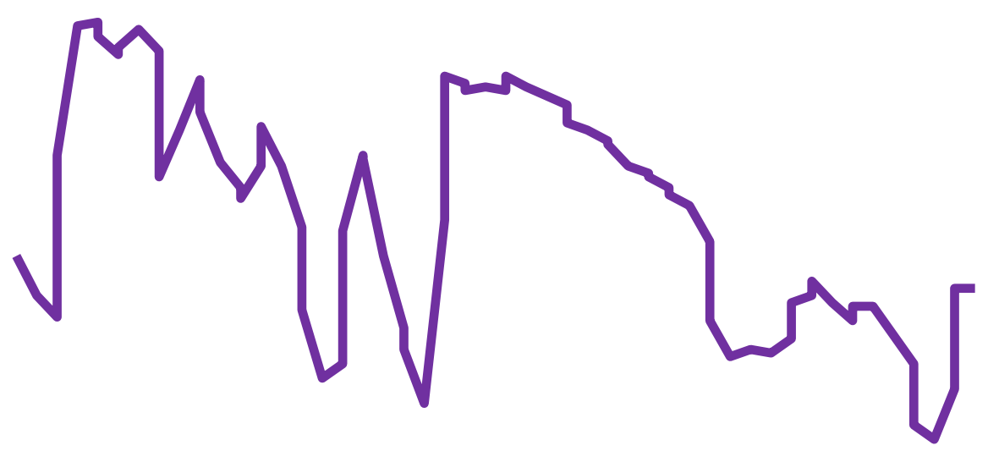

Figure 1: Visualization of a Time Series

Time series T is an ordered sequence with real number entries. It can be written as $T = \{t_1, t_2, ..., t_m\}$. The length of T can be denoted as $||T||$. $||T|| = m$. Each entry is an measurement. For sequence data, it is unlikely for two similar series to share exactly the same entries. Hence, it is more natural to compare the **shape** instead of the absolutes values of the points. Euclidean distances, which simply returns the root of the summation of squares of the point differences in the same time order, cannot do the job. Dynamic time warping, which is a technique from the speech recognition community, is proved to be superior in many domains.

Brief Review of US, DTW and USDTW

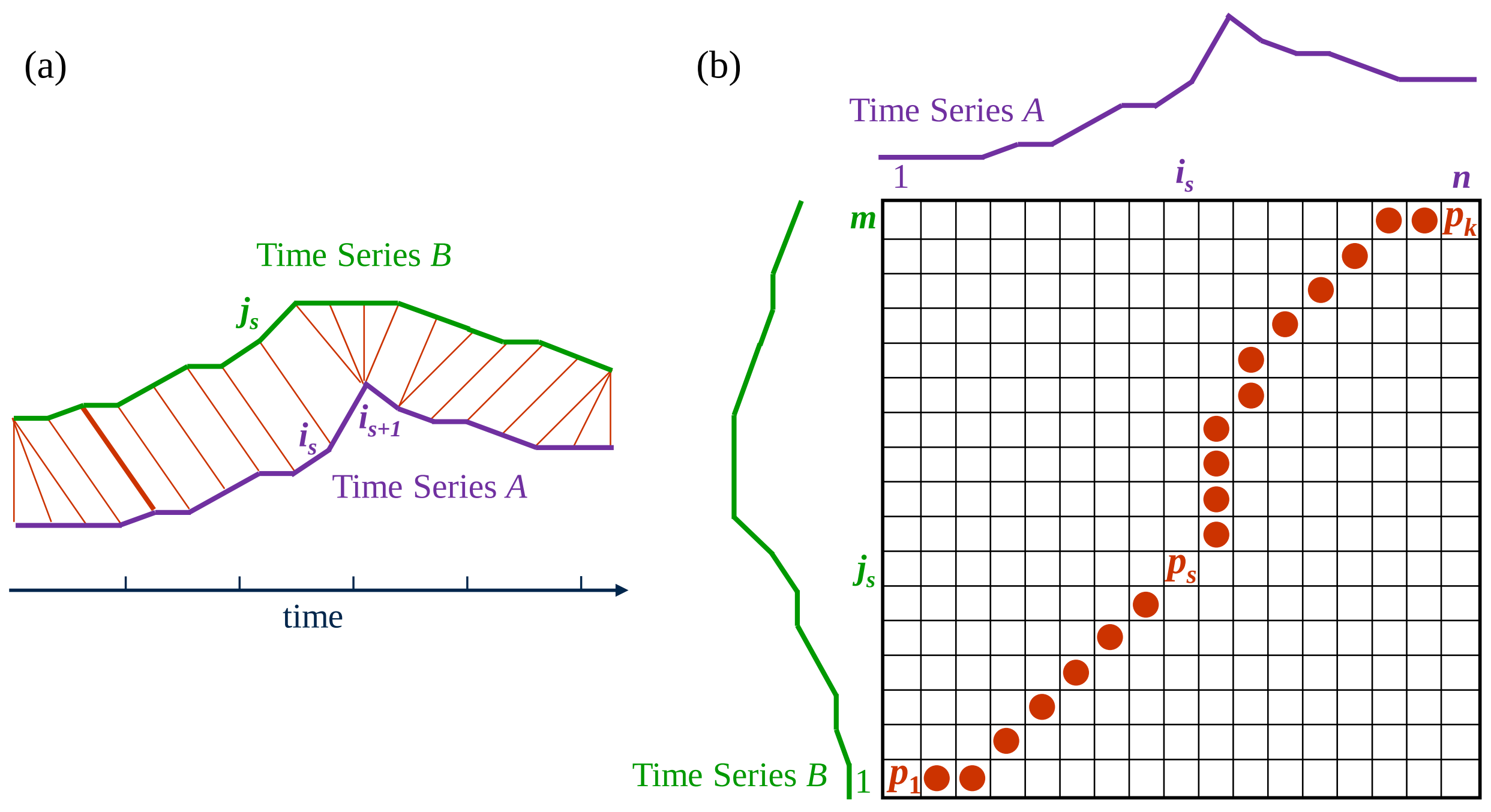

Figure 2: Dynamic Time Warping between two series

Figure 2 shows the mechanism of DTW. Given two time series, A and B, we want to compare their similarity. Dynamic time warping uses a computational technique namely "Dynamic Programming" to construct a matching matrix, which is shown in (b) in the figure. Each cell in the matrix stores the point-wise difference between two points i and j. DTW will find a monotonic (i.e., evolve in one direction), contiguous (i.e., all the cells in the path are connected) path that has the minimum of the summation of the differences contributed by the consisted cells. This path is marked as red in (b). In (a), a point namely i_{s+1} on A will map to 4 points on B. It can also be seen in the matrix that there are 4 points aligned vertically next to the cell denoted as p_s.

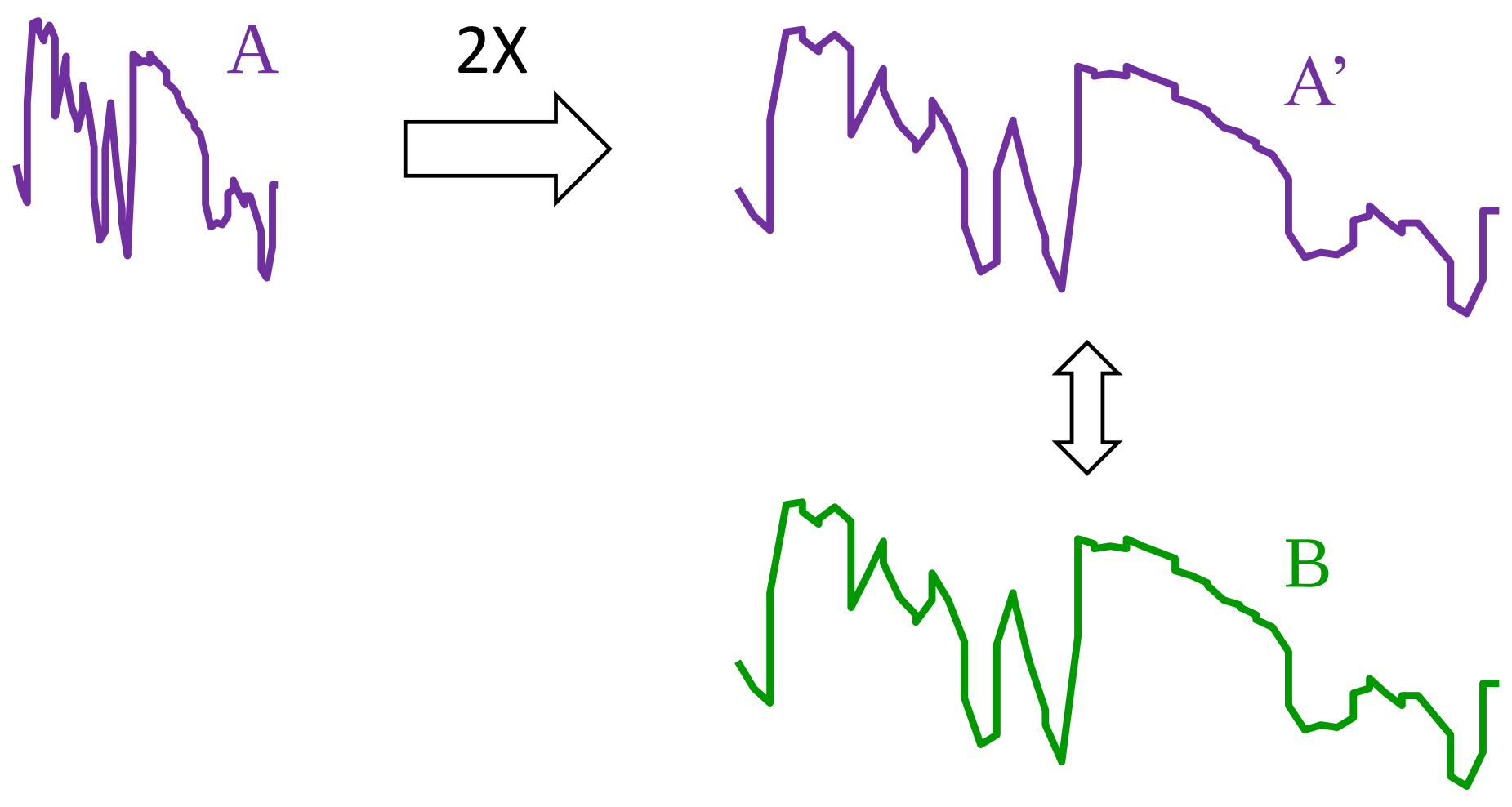

Figure 3: Uniform Scaling betwenn two sereis

If the two series have a large difference in the y-axis scale, even DTW cannot return a convincing result. Another distance measure called "Uniform Scaling" is designed to handle this. Basically, it first tries to find the scaling factor of one series with respect to the other series. Then, using this scaling factor to compress(stretch) the series so that it is in the similar scale of the other one. In Figure 3, we have two series namely A and B. They are indeed exactly the same but with different scale. Uniform scaling tries to stretch A by two time to produce another series called A'. A' will have a good comparison with B as they are in the same scale. After the scaling, DTW is applied.

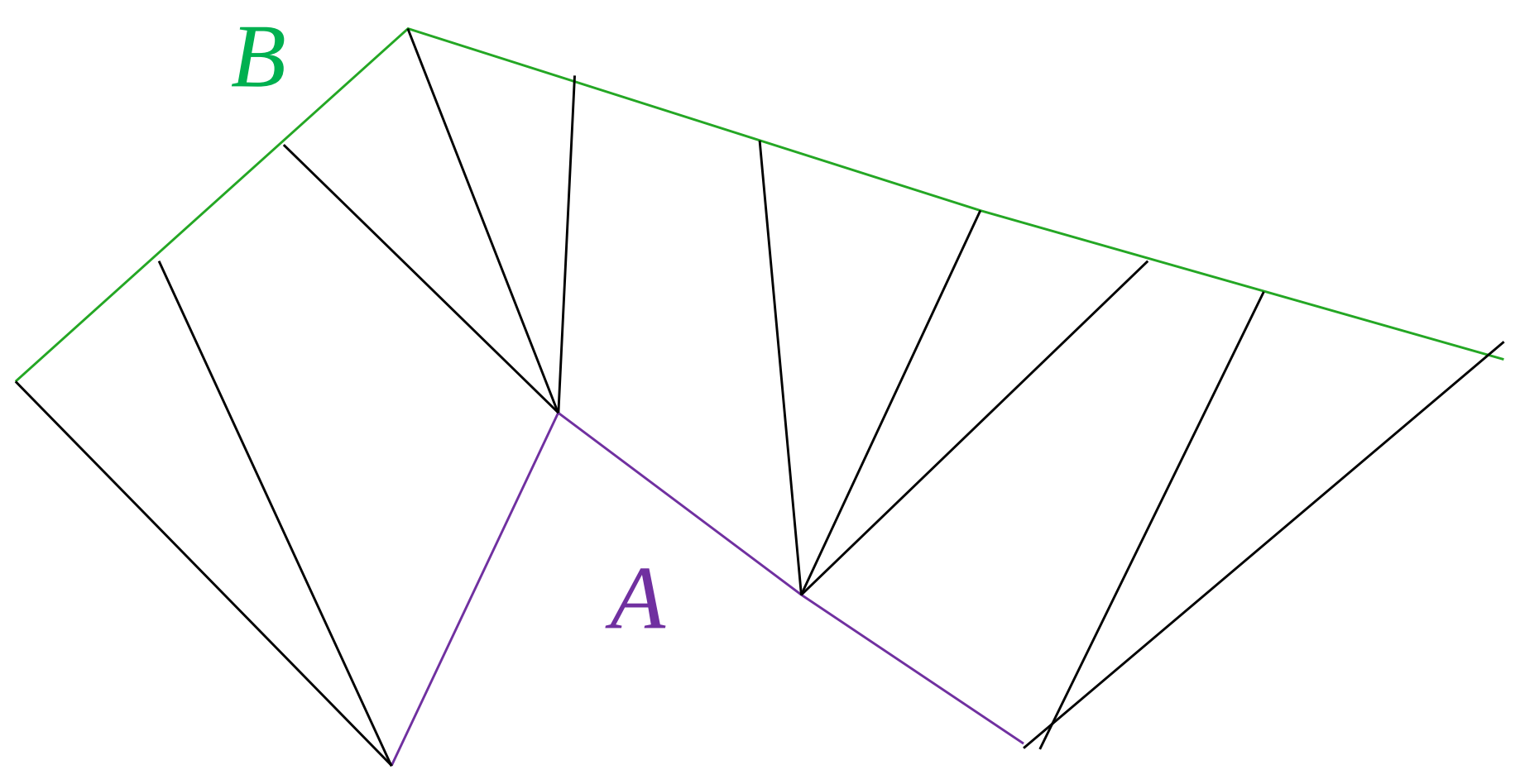

Figure 4: A toy example showing that DTW is not a generalization of US

Suppose there are two time series, $A = \{0, 3, 2, 1\}$ and $B = \{0, 1, 2, \mathbf{3}, 2.67, 2.3, \mathbf{2}, 1.6, 1.3, \mathbf{1}\}$. Indeed, we can produce B from A by stretching for 2.5X with interpolation. $A' \equiv B$.Hence, A and B would be considered as exactly the same under uniform scaling with scaling factor = 2.5X. However, DTW(A, B) $= |0-0|+|1-0|+|2-3|+|3-3|+|2.67-3|+|2.3-2|+|2-2|+|1.6-2|+|1.3-1|+|1-1| = 3.33$. This result counters our intuition.

The following sections present our novel similarity measure, namely USDTW with multiple scaling factors.

USDTW with Multiple Scaling Factors

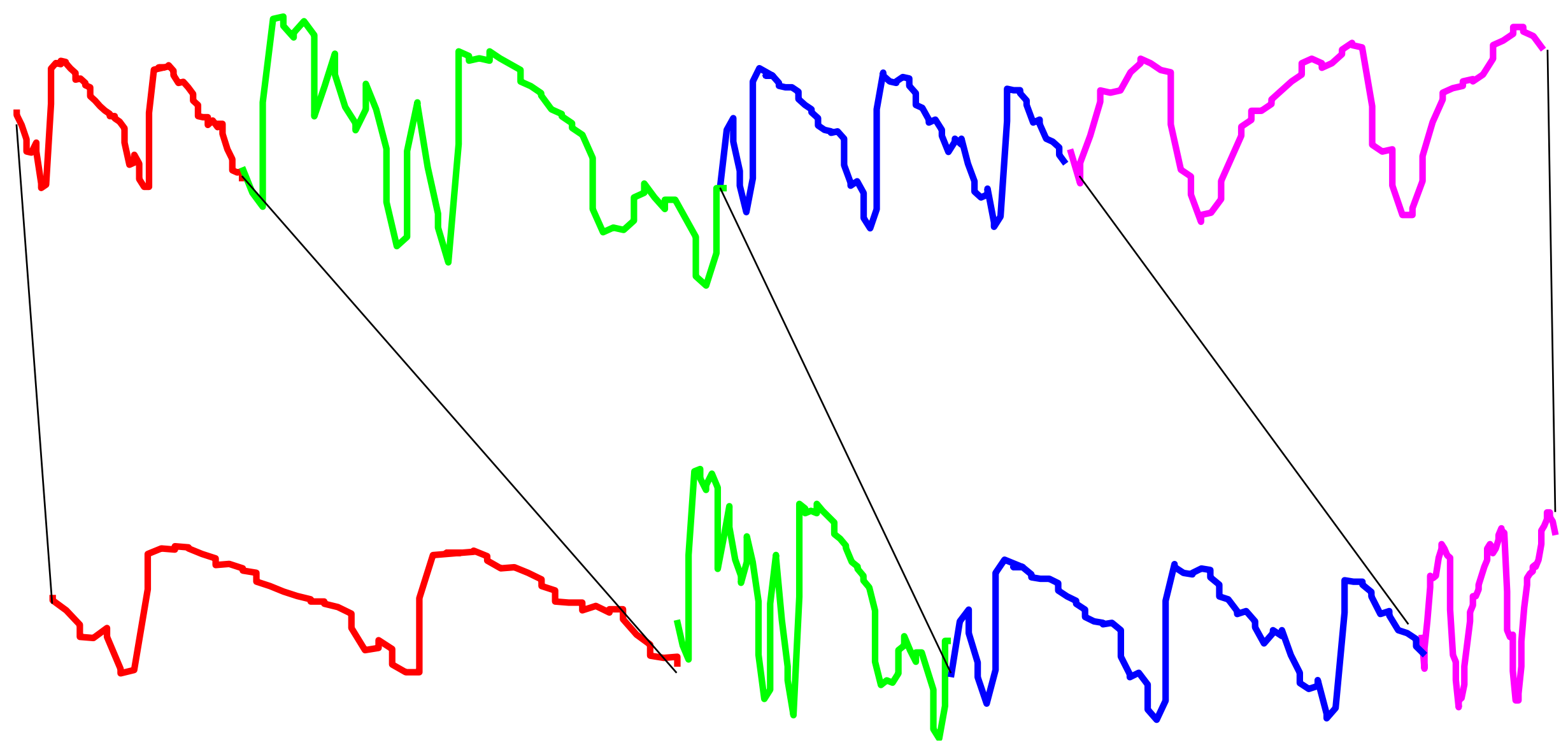

Figure 5: Motivation example showing the needs of uniform scaling with multiple scaling factors

Consider Figure 5, these two time series are made of the same set of subsequences but with different scaling factors. For example, the red component is longer in the bottom time series. This two time series could be regarded as the same in some domains in the sense that they represent the same overall mechanism. Obviously, DTW and US with one single scaling factor cannot handle this scenario. We need to have a new distance measure that can leverage multiple scaling factors to stretch(compress) the consisted components before applying DTW to measure the similarity.

Formulation of our New Similarity Measure

If we know that there are K components, the USDTW with multiple scaling factors can be defined as follows.

$$D[|i|, |j|, K] = \min \sum_{k=0}^{K-1} USDTW(i_k, i_{k+1}, j_k, j_{k+1}) \tag{1}$$

$i_1, i_2, ..., i_{K-1}$ and $j_1, j_2, ..., j_{K-1}$ are the starting(ending) points of the segments in time series i and j. The 1st attribute in $USDTW$ represents the starting point of the kth segment in time series i. The 2nd attribute in $USDTW$ represents the ending point of the kth segment in time series i.

The 1st attribute in D refers to the ending index of subsequence of i. The 2nd attribute in D refers to the ending index of subsequence of j. The 3th attribute in D refers to the number of the consisted components. $D[|i|, |j|, K]$ is our goal.

$$D[|i|, |j|, 1] = USDTW(1, |i|, 1, |j|) \tag{2}$$

$D[|i|, |j|, 1]$ is equivalent to the USDTW applying on the whole i and whole j with only one component. It is USDTW with one scaling factor.

$$D[i', j', 1] = \min_{i<i', j<j'} \{D[i-1, j-1, k-1] + USDTW(i, i', j, j')\} \tag{3}$$

By using these 3 equations, USDTW with multiple scaling factors for two sequences could be computed in recursive manner.

References

[1] John Aach and George M Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 2001.

[2] Ada Wai-Chee Fu, Eamonn Keogh, Leo Yung Lau, Chotirat Ann Ratanamahatana, and Raymond Chi-Wing Wong. Scaling and time warping in time series querying. *The VLDB JournalThe International Journal on Very Large Data Bases*, 17(4):899–921, 2008.