Pattern Matching Via Sequence Alignment: Analysing Spatio-Temporal Distances

S. Stehle

GeoVISTA Center, Department of Geography, Pennsylvania State University
814-865-3433
samstehle@psu.edu

1. Introduction

This paper presents sequence alignment as a novel analytic method for analyzing patterns. The social and earth sciences devote considerable research to understanding patterns. Discovering theoretically driven patterns dominates this research, however the process of pattern matching links those patterns to sequences of real events. In the last five decades, research that computationally identifies and compares patterns in space and time investigates human and animal movement (Hagerstrand 1970; Laube, Imfeld, and Weibel 2005), weather tracking (Beard, Deese, and Pettigrew 2007), and the spread of disease (Waller et al. 1997), among others. However, spatio-temporal dynamics are not limited to movement vectors. Geography has also considered sequences in a spatio-temporal context that do not explicitly contain references to moving entities, such as the diffusion of innovations (Hagerstrand 1967). This paper examines spatio-temporal processes through the patterns that emerge from seemingly disconnected events in space and time. It reconsiders sequence alignment as a tool to expand current geographic research on patterns of events. This paper extends the sequence alignment method, an algorithm developed in the biological sciences, by specifically focusing on patterns of events that indicate political and economic transition.

2. Methods

Needleman and Wunsch (1970) designed the sequence alignment algorithm for use on sequences without temporal properties. Thus, this paper introduces several modifications to analyze temporal event sequences. These modifications reflect existing tools that incorporate sequence alignment, especially the Clustal series and ClustalG (Wilson, Harvey, and Thompson 1999) in particular. The expanded tool was created using Java.

2.1 The Sequence Alignment Procedure

The alignment process is based on a technique of changing two compared sequences so that patterns approximate one another as closely as possible. This involves inserting "gaps" into each sequence to make equivalent elements in both sequences occur in corresponding positions. The placement of gaps within compared sequences is important to generating an optimal alignment, through which the number of inserted gaps is minimal.

Alignment of two sequences is achieved using a two-dimensional similarity matrix with the x- and y-axes composed of the elements in both sequences. Each cell in the matrix is given by the sum of a binary that quantifies the elements' equivalence plus the highest value in the next adjacent row and column. The matrix is then traversed to maximize the sum of "collected" cell

values along the way (Needleman and Wunsch 1970). Traversing the matrix yields a single optimal combination of elements at corresponding positions from both sequences. Gaps are placed within a sequence when the traversal skips an adjacent row or column. This indicates that the sequence must be extended to accommodate for unexpected elements in one of the sequences. Gaps measure the temporal distance between compared sequences.

2.2 Temporally-Optimal Alignment

The few applications of sequence alignment that incorporate time use the method to analyze sequences whose elements are representative of successive time intervals. This paper accounts for event sequences that contain multiple events per unit of time. The result of an event-based conceptualization is a multi-dimensional sequential representation of time, which complicates the single-dimensional abilities of existing sequence alignment procedures.

The original procedure assumes an optimal alignment is one that traverses the matrix directly diagonal with no skipped cells. However, because event sequences may contain multiple events per time step, the tool accommodates skipped cells while still maintaining sequential structure. This is accomplished by a modification of the algorithm that aligns sequences based on the time stamp of each event element, not its index within the sequence.

2.3 Modified Edit Distance

The use of gaps changes the composition of aligned sequences. Thus, the insertion of gaps produces temporal significance. The edit distance, which is a sum of added gaps, is therefore a unit of additive time, where insertions and deletions of elements become insertions and deletions of time. An edit distance with a label makes sequence alignment useful as a standalone tool that measures the temporal difference in the ways that two patterns transpire.

2.4 Tool Appearance

This tool has four parts that appear as separate displays (see figure 1). The first is a sequence view that displays the original pair of sequences. Next is a matrix view that displays the similarity matrix specifically for debugging. Perhaps the most important display is an alignment view that displays the resulting alignment with gaps following the analysis. This view provides visual cues to explore correspondence between event pairs. Finally, an edit distance view displays the calculated temporal distance between the sequences.

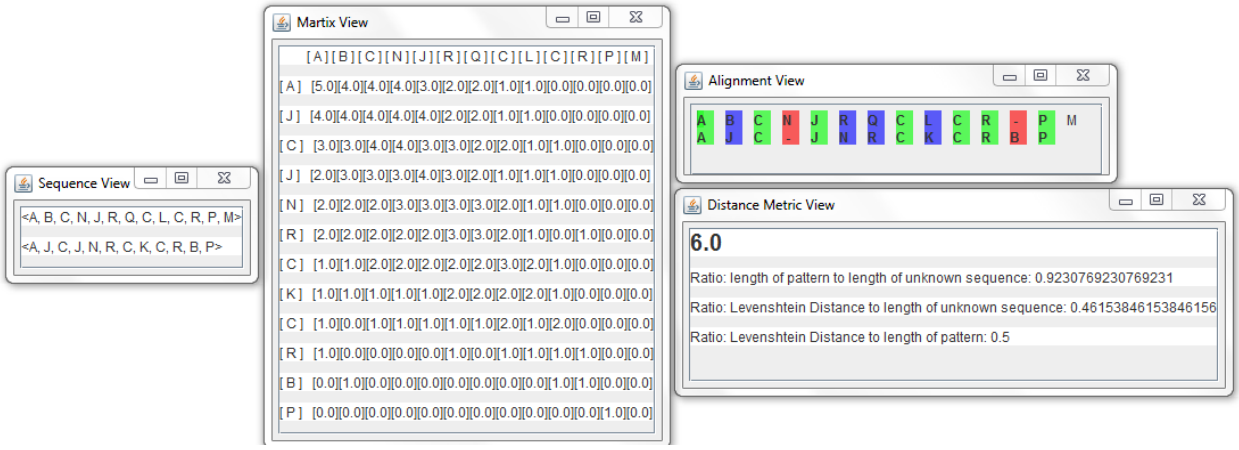

fig. 1. Four views of the sequence alignment tool demonstrated on sample character sequences

3. Pattern Analysis of Political Events

This modified sequence alignment method utilizes patterns derived from the STempo environment for pattern discovery. Patterns are defined by a sequence of events, each of which follows the previous event within a critical interval of time. Patterns that exemplify processes of political transition in Yemen in 2011/2012 have been identified by the STempo environment and compared to events captured during previous political turmoil in the country in the early 1990s. Patterns consisting of similar events were found to transpire in slightly different, but not unexpected ways. The temporal edit distance shows a significantly increased amount of time between similar events during the unification of North Yemen and South Yemen in the early 1990s compared to the expected patterns identified during the Arab Spring of early 2011. The increased amount of time identified by the temporal edit distance was then explored in the alignment view of this tool. The alignment shows that the increased time present in the earlier patterns was distributed evenly between all pairs of events. This indicates the need for further research to examine the role of access to technological advances in communication among other variables through classification of events.

4. References

Beard, K., H. Deese, and N. R. Pettigrew. 2007. A Framework for Visualization and Exploration of Events. *Information Visualization* 7:133-151.

Hagerstrand, T. 1967. *Innovation Diffusion as a Spatial Process*.(Allan Pred and Greta Haag trans.). Chicago: University of Chicago Press. (Original edition published 1953).

———. 1970. What About People in Regional Science? *Papers in Regional Science* 24:7-24.

Laube, P., S. Imfeld, and R. Weibel. 2005. Discovering Relative Motion Patterns in Groups of Moving Point Objects. *International Journal of Geographical Information Science* 19:639-668.

Needleman, S. B., and C. D. Wunsch. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology* 48:443-453.

Waller, L. A., B. P. Carlin, H. Xia, and A. E. Gelfand. 1997. Hierarchical Spatio-Temporal Mapping of Disease Rates. *Journal of the American Statistical Association* 92:607-617.

Wilson, C., A. Harvey, and J. Thompson. 1999. ClustalG: Software for Analysis of Activities and Sequential Events. Paper presented at Workshop on Longitudinal Research in Social Science: A Canadian Focus, London, Ontario, CA, 25-27 October 1999