Similarity maps of behavior sequences: Methods and software for pattern exploration and segmentation

Vicenç Quera
Department of Behavioral Science Methods, University of Barcelona, Spain, vquera@ub.edu

Researchers who study interaction are usually interested in detecting temporal patterns within observed sequences of behavioral codes. A pattern is "the repeated or regular way in which something happens or is done" (Collins COBUILD English dictionary); pattern is then synonymous with order and predictability, and the opposite of randomness and noise. Therefore, searching for patterns in sequences of codes that represent interaction amounts to detecting whether two or more codes often occur in succession, whether they tend to occur at the same time, or whether they tend to occur within a specific time intervals of each other. Different types of patterns can be potentially detected depending on how interaction codes are represented (as event sequences, timed event sequences, etc.; for a classification of types of sequential data, see [1]). A classical, quantitative approach to describing the sequence (e.g., [2]) consists in tallying transition frequencies among the codes and computing inferential statistics in order to know whether some transitions tend to occur more often than expected in case of randomness. However, besides that kind of molecular result, an exploration of the sequence as a whole can provide new insights as to whether certain patterns exist, where in the sequence they occur, and even whether they tend to repeat in different but comparable sequences. Such global exploration can be carried out by first computing indices of similarity among parts of the sequence (within specified time windows), then representing them as a two-dimensional map. Code repetitions, chains composed of certain codes occurring in succession, and possible patterns consisting of codes separated by a more or less constant interval may be detected by visual inspection of that map. Moreover, the map as a whole can indicate whether the sequence is probably random or whether it contains patterns, and, if so, where they tend to be located in the sequence.

Program RAP (for RAndom Projection, [3]) transforms event or timed event behavior sequences into such a map; both intra- and between sequences similarities can be computed. The program represents successive intervals or time windows in the sequence by quantitative vectors by means of an analytical technique called "random projection" [4], then computes the similarity between every time window and every other time window. Similarities, which are values ranging from 0 (no similarity between the two windows) to 1 (perfect similarity), are then displayed as grey pixels in an image, the greater the similarity the darker the pixel. Researchers can then visually inspect the image, and navigate it with the mouse cursor on a computer screen in order to explore its regions. These graphical representations are analogous to dot-plots (for exploring self-similarities in texts and in sequences of nucleic acids; e.g., [5], [6]), recurrence plots (for the analysis of continuous time series describing the behavior of dynamical systems; e.g., [7]), and waveform similarity plots (for the analysis of structure and classification of digital media streams; e.g., [8]).

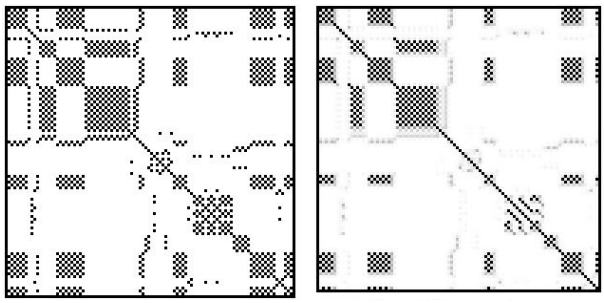

Figure 1. *Self-similarity maps created by program RAP for an event sequence of verbal interaction.*

Figure 1 shows two self-similarity maps created by RAP for an event sequence of verbal interaction in a couple. For the two maps, the sequence is represented top to bottom, and left to right; the map on the left was obtained by applying a moving time window containing one only code, and the one

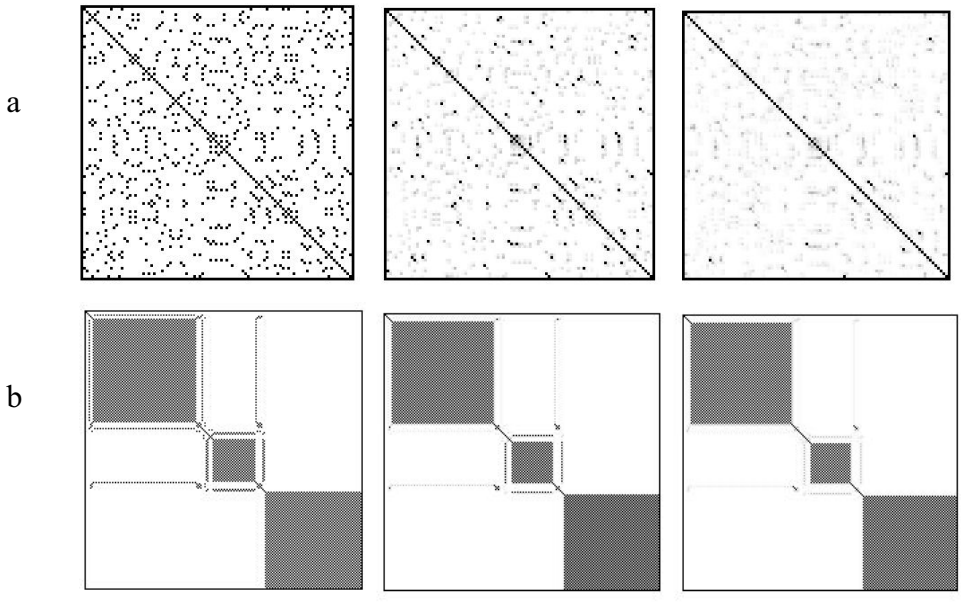

Figure 2. *Self-similarity maps for (a) a random event sequence, and (b) a highly patterned event sequence or verbal interaction.*

on the right was obtained by applying a window containing three consecutive codes (in the latter case successive windows overlapped). In both cases, main diagonals represent the similarity of every time window with itself, which is obviously maximum, and the images are symmetrical around their diagonals. Chequered regions indicate alternation of windows containing identical or very similar data; diagonal segments in the images, parallel to the main diagonal, indicate that certain chains of similar windows repeat in different parts of the sequence. Other hints about possible patterns are vertical and horizontal lines, either continuous or fragmented, which indicate that a certain code, or chain of codes, occurring in a position in the sequence repeats in several other positions. As a whole, a similarity map reveals general features of the sequences, and can be useful for classifying the interaction as patterned or random.

Figure 2a shows three self-similarity maps (for moving time windows containing 1, 2, and 3 codes, respectively) for an event sequence of verbal couple interaction in which husband responds to wife at random, and vice versa; Figure 2b shows three self-similarity maps for an event sequence of verbal interaction containing long runs of reciprocal interactions, which correspond to three big chequered squares along the main diagonal. For a random sequence, dark pixels in the similarity map tend to vanish rapidly as the width of the moving time window increases, while for a highly patterned sequence the proportion of dark pixels remains more stable. A self-similarity map can be further processed in order to reveal the temporal structure of the sequence; program RAP can detect segments in a sequence by correlating a Gaussian checkerboard filter along the map's main diagonal, and computing a novelty score (a technique proposed by [9]). Figure 3 shows a self-similarity map for a timed event sequence of mother-child interaction, obtained by applying a moving time window 20 s wide (successive windows overlapped by 10 s), and the resulting novelty score, whose peaks indicate segment boundaries, or temporal points at which significant changes are detected in the behavioral sequence.

Program RAP runs on Windows systems, reads sequence data files in *Sequential Data Interchange Standard* (SDIS) format (see [1]), and saves similarity maps and related graphical information as BMP images included in HTML documents. The program can be downloaded from www.ub.es/comporta/vquera

References

1. Bakeman, R., & Quera, V. (1995). *Analyzing interaction. Sequential analysis with SDIS and GSEQ*. New York: Cambridge University Press.
2. Bakeman, R., & Quera, V. (1995). Log-linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin* **118,** 272-284.
3. Quera, V. (2008). RAP: A computer program for exploring similarities in behavior sequences using random projections. *Behavior Research Methods* **40,** 21-32.
4. Mannila, H. & Seppänen, J. (2001). Recognizing similar situations from event sequences. *Proceedings of the First SIAM Conference on Data Mining*, Chicago. Available at: http://www.cs.helsinki.fi/~mannila/postscripts/mannilaseppanensiam.pdf
5. Church, K.W., & Helfman, J.I. (1993). Dotplot: A program for exploring self-similarity in millions of lines of text and code. *Journal of Computational and Graphical Statistics* **2**, 153-174.
6. Maizel, J.V., & Lenk, R.P. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proceedings of the National Academy of Sciences* **78,** 7665-7669.
7. Eckmann, J.P, Kamphorst, S.O., & Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhysics Letters* **5,** 973-977.
8. Cooper, M., & Foote, J. (2002). Automatic music summarization via similarity analysis *Proceedings of the International Symposium on Music Information Retrieval*, 81-85.
9. Foote, J., & Cooper, M. (2003). Media segmentation using self-similarity decomposition. *Proceedings of SPIE,* **5021,** 167-175.

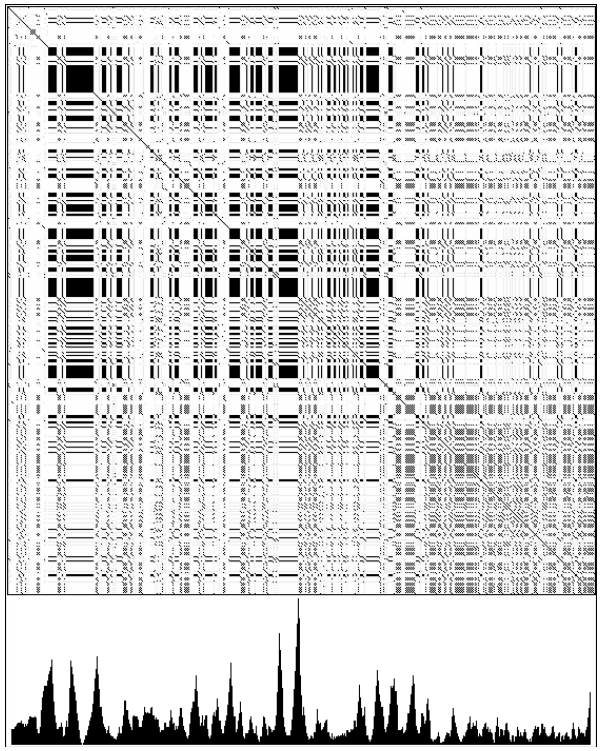

Figure 3. *Self-similarity map for a timed event sequence of mother-child interaction, and novelty scores indicating segment boundaries in the sequence.*