Harpoon or maggot?
A comparison of various metrics to fish for sequence patterns

Nicolas Robette, Printemps (UVSQ-CNRS, UMR 8085)
Xavier Bry, I3M, Université de Montpellier II

More often than not, the first step of holistic approaches consists in measuring the dissimilarity between life courses (regarded as sequences). Pairwise distances between sequences can further be used in various ways, often with data reduction techniques such as multidimensional scaling or clustering. Many dissimilarity metrics exist in various domains (bioinformatics, data mining...) and their use in social sciences has been developing rapidly for a decade or two. The most widely known is certainly Optimal Matching Analysis (Abbott & Forrest, 1986), but other metrics for sequence analysis have been proposed and similar techniques using correspondence analysis also exist. Therefore, a crucial and pervasive issue in papers using holistic approaches is robustness: to what extent do the various techniques lead to consistent and converging results? What kinds of patterns does each of the metrics identify most effectively?

Numerous articles have been devoted to comparing metrics. However, most of them have limitations: they deal with a narrow range of methods at a time; they apply to specific sets of empirical data; other choices implied in the holistic approach (clustering techniques, etc.) may blur the results. So generalization is often problematic. We propose a systematic comparison of a collection of metrics that have been used in the social science literature, based on the examination of dissimilarity matrices computed from two data sets: a simulated one comprising various sequence patterns that sociologists may aim at identifying, and an empirical one (about occupational careers) as a "control sample". Thus what we are trying to do here is not to point out a hypothetical "best metric", but rather to unravel the specific patterns to which each alternative is actually more sensitive.

We will successively present a short review of existing methods for sequence analysis, a summary of the comparisons conducted in the literature, our own protocol for comparison, and finally, our results and discussion.