How Much Does It Cost?: Optimization of Costs in Sequence Analysis of Social Science Data

Jacques-Antoine Gauthier, Eric D. Widmer, Philipp Bucher and Cédric Notredame

Sociological Methods Research 2009; 38; 197
DOI: 10.1177/0049124109342065

Sociological Methods & Research
Volume 38 Number 1
August 2009 197-231

10.1177/0049124109342065
http://smr.sagepub.com
hosted at
http://online.sagepub.com

How Much Does It Cost?

Optimization of Costs in Sequence Analysis of Social Science Data

Jacques-Antoine Gauthier
University of Lausanne, Switzerland
Eric D. Widmer
University of Geneva, Switzerland
Philipp Bucher
Swiss Institute of Bioinformatics and Swiss Institute for Experimental Cancer Research, Lausanne Switzerland
Cédric Notredame
Centre National de la Recherche Scientifique, Marseille, France, and Centre for Genomic Regulation, Barcelona, Spain

One major methodological problem in analysis of sequence data is the determination of costs from which distances between sequences are derived. Although this problem is currently not optimally dealt with in the social sciences, it has some similarity with problems that have been solved in bioinformatics for three decades. In this article, the authors propose an optimization of substitution and deletion/insertion costs based on computational methods. The authors provide an empirical way of determining costs for cases, frequent in the social sciences, in which theory does not clearly promote one cost scheme over another. Using three distinct data sets, the authors tested the distances and cluster solutions produced by the new cost scheme in comparison with solutions based on cost schemes associated with other research strategies. The proposed method performs well compared with other cost-setting strategies, while it alleviates the justification problem of cost schemes.

Keywords: *sequence analysis; optimal matching; trajectories; empirical cost optimization*

Optimal matching analysis (OMA) has emerged since the 1990s as a main methodological innovation in the social sciences for finding

Authors' Note: Please address correspondence to Jacques-Antoine Gauthier, University of Lausanne, SSP–MISC, Bâtiment de Vidy, CH-1015 Lausanne, Switzerland; e-mail: Jacques-Antoine.Gauthier@unil.ch.

patterns in sequences of social events (Abbott and Tsay 2000). It is based on the assumption that succession of social statuses or events constitutes stories throughout the life course that can be measured in a set of data (Abbott 1984, 1990a, 1990b, 1995a, 2001). Usual measures of distance, such as the Euclidean distance, are ineffective for many sequential data, for example, when their lengths differ (Kruskal 1983; Abbott 1995b, 2001). Therefore, multivariate statistical methods falling within the framework of dynamic programming procedures and stemming from molecular biology (e.g., Needleman and Wunsch 1970) have been adapted to the study of social trajectories (Abbott and Hrycak 1990; Erzberger and Prein 1997; Giele and Elder 1998; Wilson 1998; Aisenbrey 2000, Rohwer and Pötter 2002) and embodied in various softwares (TDA,[1] Optimize,[2] and CLUSTALG[3]).

One problem identified as major in this set of methods, however, lies in the cost schemes on which empirical analyses are based. As a matter of fact, optimal matching methods decompose the total difference between any two sequences into a collection of individual elementary differences using substitution, deletion, and insertion operations (Kruskal 1983). The determination of the costs attributed to those operations is the subject of an ongoing debate in the social sciences (Abbott and Tsay 2000; Wu 2000) since the setting of costs is not in most cases based on explicit and strong theoretical stances. For example, given a pair of sequences to be aligned, one can wonder if it is the same to substitute 1 year of full-time employment with either 1 year of part-time employment or 1 year of being exclusively at home. If it is not, we should consider weighting the costs of those operations so that they contribute differently to the final alignment of the two sequences. Some scholars emphasize the large impact that cost setting has on the final results of their analysis (Rohwer and Pötter 2002) whereas others take the opposite stance, underlining its marginal impact on similarity scores among sequences (Levine 2000). However, most argue for both sensitivity and stability of the effect of cost variations on the results of the analysis (Abbott and Hrycak 1990).[4] Therefore, researchers in the social sciences are left wondering to what extent the final results of their analyses are reproducible and valid. This article first describes usual solutions proposed by social scientists in regard to the problem of the determination of costs in sequence analysis. Then, it proposes a method that computationally derives costs from the empirical data, based on state-of-the-art approaches in bioinformatics (Henikoff and Henikoff 1992; Müller and Vingron 2000; Ng, Henikoff, and Henikoff 2000; Yu and Altschul 2005). The proposed algorithm is then tested on three distinct social science data sets. We further discuss the consequences of the results for empirical analyses of sequence data in the social sciences.

How Are Costs Determined in the Social Sciences?

The issue of costs concerns two operations in sequence analysis: substitution and insertion/deletion. Because this stage of sequence analysis is critical for further results, all publications that use OMA provide some sense of how costs are set, but with unequal degrees of details. Based on a literature review in the field, we found five strategies regarding the setting of substitution costs as they are used in the social sciences.

A first strategy is to set all substitution costs to a constant, that is, using an identity matrix (Dijkstra and Taris 1995; Rohwer and Trappe 1997; Pentland et al. 1998; Wilson 1998; Schaeper 1999; Billari 2001). Those using this strategy argue that they have no rational way to set costs in another way. This strategy is used largely when no theoretical rationale is available for supporting the setting of costs. It has been criticized, however, for its inability to reflect unequal differences between a given set of social characteristics, on one hand, and the distribution of those different positions on the other. Abbott and Hrycak (1990) gave the example of determining the proximity of some occupational positions such as senior executive, first-level supervisor, and line worker, which would be impossible. They proposed that in this case substituting or inserting the rarest one should be more costly.

A second research strategy uses differentiated costs following theoretical intuitions concerning the "social weight" for substituting one status with another (Chan 1995; Erzberger and Prein 1997; Halpin and Chan 1998; Blair-Loy 1999; Giuffre 1999; Schaeper 1999; Scherer 2001; Widmer, Levy, et al. 2003). For instance, Chan (1995) underlined that decisions about costs have to be grounded in theoretically important divisions between social classes. One may agree only in principle with this and comparable statements, but the social sciences are currently characterized by various contradicting theories rather than by a common theoretical framework such as evolutionary theory in biology (Grauer and Li 2000; Turner 2001; Giddens, Duneier, and Applebaum 2003). Therefore, backing costs with theoretical statements often proves difficult because of the large number of alternatives, depending on the theory chosen. Also, because results from sequence analysis are used to support and contradict theoretical statements at the same time, there is some circularity in building the costs on the same theoretical statements that they are supposed to help prove or disprove. This is as true for research on social classes as for other research areas in the social sciences.

A third strategy consists of applying some empirical coding scheme based on common sense or face value. Aisenbrey (2000) set the substitution costs

according to a hierarchical ordering of the statuses that constitute the sequences. Abbott and Forrest (1986, 1989), for instance, categorized the statuses of sequences according to the number of steps up the hierarchy necessary to put them under a common heading. The substitution cost is computed as the ratio of this number to the total number of steps possible. Applying the "garbage can model" to estimate the institutional influence on the textbook publishing process in physics and sociology, by means of sequence comparison technique, Levitt and Nass (1989) based the setting of their substitution costs on a list of topics and subtopics used in structuring textbooks. The cost was set to 1 for a change from one topic to another (e.g., stratification vs. ideology) and to 0.5 for a change between subtopics of the same topic (race vs. gender as substructures of stratification). Studying the structure of sociological articles across time, Abbott and Barman (1997) defined two levels of elementary states of sociological articles. Level 1 comprised statuses such as "introductory," "hypotheses," and "literature," whereas Level 2 encompassed subdivisions such as "topic," "state of affairs," "questions," and "author's theory/assertion" for the introductory heading. A substitution cost of 1 was attributed to subheadings falling under different headings and 0.25 for subheadings falling under the same heading. In all cases reviewed, the setting of costs is not done on strong theoretical bases, but rather on rules that make empirical sense considering the problem at hand.

Fourth, some authors set costs based on a combination of common sense (the third strategy) with the likelihood of transitions between statuses in the empirical data (Abbott and Hrycak 1990; Stovel, Savage, and Bearman 1996; Stovel and Bolan 2004). For instance, in their programmatic study of musicians' careers, Abbott and Hrycak (1990) first distinguished for each musician nine spheres of activity (court, town, church, etc.) and 15 positions (vocalist, composer, Kapellmeister, etc.). Among the 135 combinations, they finally kept 35 different occupational positions as statuses in a musician's career. To set the costs of substitution, they proposed that a change in both sphere and position is more drastic than a change in only one sphere. They set to 0.75 the cost for a change within either a sphere or a position. The cost was set to 1 when the change occurred on both levels. Second, in order to take into account the fact that some pairs of occupational positions seem to be closely connected with mobility (i.e., they often lie on the same career line), they combined the distance matrix, based on mobility, with a position/sphere dissimilarity matrix. This matrix was constructed by classifying all moves in all careers according to their frequency. The final substitution matrix is then a linear combination of corresponding symbols of the two matrices.

An alternative to the development of substitution costs is represented by the use of transition costs, estimated directly on collections of trajectories. Such an option is available in TDA software, but to our knowledge, no empirical results based solely on this way of determining substitution costs have yet been published. In a transition matrix, low costs indicate pairs of symbols that are likely to co-occur in a specific life trajectory (such as work and retirement). In a substitution matrix, low substitution costs indicate symbols that are likely to occur simultaneously in two different trajectories. A low substitution cost does not imply any transition, but rather an equivalence of some sort between the two considered statuses. While transition matrices are ideal for analyzing individual strings and identifying trajectory anomalies, they are much less suitable to comparisons of alternative trajectories that rely on the comparison of symbols occurring simultaneously in different trajectories.

Some scholars have used costs based on transitions, combined with some additional criteria. Stovel et al. (1996) derived the substitution costs from an analysis of the complete transition matrix reporting the distribution of work transitions of all workers of Lloyds Bank over the period 1890 through 1970. They then distinguished costs for positions and for branch changes and combined them. Considering residential trajectories, Stovel and Bolan (2004) used a similar strategy. They first constructed a place-type variable (nine categories) based on a continuum ranging from small rural towns to large metropolitan cities. This theoretically based distinction was then combined with the empirical distribution of the frequency of all possible transitions among types of places. The substitution matrix was constructed as a repeated adjustment between the initial theoretical model and the empirical transition rates.

In contrast to the previous three strategies, this strategy marks a significant improvement as it is at least partially empirically driven. There are, however, various problems existing with the solutions currently proposed. First, all reviewed solutions are at least partially driven by intuition or face value, or by some kind of theoretical stance. Second, the choice of simple frequencies (or a linear function of them) to weight the substitution cost is not supported by any formalized computational methods nor by any statistical theoretical grounds. Third, even in cases where ''pseudo'' or intuitive iterative methods are used to set the substitution costs (cf. Stovel and Bolan 2004), no formal rules are presented that justify the solution chosen by the researchers. Fourth, none of those models succeed in giving a systematic and fully empirically driven procedure of substitution cost settings. Finally, no attempts are made to optimize costs based on the empirical data at hand.

In the fifth strategy, some researchers acknowledge having used a mix of several if not all approaches listed above, insisting on the exploratory dimension of the process and the fact that guidelines are few and rather fuzzy (Rohwer and Pötter 2002). To summarize, researchers in the field have underlined that the issue of the determination of costs in OMA remains presently open.

An Alternative: Deriving the Cost Empirically

To develop a more systematic and reliable method for cost setting to the ones currently existing in the social sciences, one should get back to the basics of sequence alignment. Given two strings I and J, Given two strings I and J, a penalty for insertions and deletions (called INDEL), and a cost matrix C, where $C_{S_i S_j}$ is the cost for aligning S_i, the ith symbol of I, against S_j, the jth symbol of J, the score of the optimal alignment can be computed using the following recursion:

$$OMA(i,j) = Best \begin{cases} OMA(i-1,j-1) + C_{S_i S_j} \\ OMA(i-1,j) + INDEL \\ OMA(i,j-1) + INDEL \end{cases}. \quad (1)$$

In a general sequence comparison perspective, one considers that a substitution is equivalent to a deletion followed by an insertion. Therefore, the value of an INDEL is often arbitrarily set to half of that of a substitution (Kruskal 1983). Each line in equation (1) corresponds to the optimal match score of two substrings. For instance, $OMA(i-1,\ j-1)$ corresponds to the optimal match score of a subsequence containing the symbols 1 to $i-1$ of Sequence 1, against a subsequence containing the symbols 1 to $j-1$ in the second sequence. As such, this equation defines a recursion in which the score of any alignment $OMA(i, j)$ can be estimated by considering an optimal extension of the three shorter alignments $OMA(i-1,\ j)$, OMA$(i-1,\ j-1)$, and $OMA(i,\ j-1)$. Considering that each of these shorter alignments is already an optimal matching of the associated substrings, we know that $OMA(i, j)$ is optimal. This strategy relies on the assumption that each position is independent and that the alignment scores are additive.

The alignment of Sequences A and B in Figure 1 is produced by applying recursion, as in equation (1), and iteratively filling up the $OMA(i, j)$ array until the optimal matching score $OMA(I, J)$ is obtained (Kruskal 1983). By recording the results of all the comparisons made at each step of the recursion, it is possible to trace back the optimal scores from

Figure 1
Example of Optimal Matching Score Computation and Alignment

SeqB: CRBEF
SeqA: ACDBDEF

OMA(i,j) array used to compute optimal matching score and alignment

			1	2	3	. . .	I
		SeqB	C	R	B	E	F
	SeqA	**0**	-1	-2	-3	-4	-5
1	A	**-1**	-2	-3	-4	-5	-6
2	C	-2	**-1**	-3	-5	-6	-7
3	D	-3	-3	**-3**	-5	-7	-8
. . .	B	-4	-5	-5	**-3**	-5	-7
. . .	D	-5	-6	-7	**-5**	-5	-7
. . .	E	-6	-7	-8	-7	**-5**	-7
J	F	-7	-8	-9	-9	-7	**-5**

Displaying optimal alignment
B: -CRB-EF
A: ACDBDEF

the cell $OMA(I, J)$, thus generating an alignment, as shown in Figure 1, where an identity substitution matrix has been used. Such a matrix assigns the value 0 to the matching of two identical letters and the value -2 to the substitution of two different letters. Insertion or deletion occurring at one extremity of the alignment takes the value -1 (terminal INDEL) and the value -2 when they are used within the alignment (internal INDEL). In the $OMA(i, j)$ array below, the traceback is indicated in bold. Starting from the bottom-right corner of the array, vertical moves correspond to an INDEL in Sequence A, horizontal moves to an INDEL in Sequence B, and diagonals to a match or a substitution.

One of the main issues that arises in equation (1) concerns the estimation of the substitution costs (C_{ij}). This issue is also central in biology. Given 20 amino acids, some so similar that they are almost interchangeable while others are very different, one cannot simply use any a priori substitution matrix; some modeling is required. Dayhoff, Schwartz, and Orcutt (1978) addressed this problem in the 1970s using a data-driven empirical approach. They manually aligned sets of highly similar, same-length sequences of amino acids and counted the number of mutations tolerated by evolution. A mutation is characterized by the presence of 2 different amino acids at the same position of the alignment. In a general sequence comparison perspective, this is called a substitution (Kruskal 1983). In this context, highly similar sequences are defined as those having more than 80 percent identity, where the percentage of identity is calculated by dividing the number of positions in the alignment in which the same letter appears in both sequences (identities) by the length of the alignment, as shown in equation (2). All positions with a gap in either sequence are nonidentities; thus, only the alignment of two identical sequences yields to 100 percent identity:

$$Percentage\ of\ Identity = W = \frac{Number\ of\ Identical\ Matches}{Alignment\ Length}. \qquad (2)$$

Selecting sequences with a high percentage of identity for computing data-driven costs of substitution prevents biases due to uncontrolled heterogeneity. For instance, the alignment of Figure 1 displays 57 percent identity (four identical pairs of letters found over the seven positions of the alignment). Finally, for each pairwise alignment, the relative frequency of substitutions occurring between two particular amino acids is compared to what was expected by chance alone. These values are computed as log odds and tabulated into a data-driven substitution matrix, as in equation (3):

$$Dayhoff\ Cost(a,b) = \log\left(\frac{f_{ab}}{f_a * f_b}\right). \quad (3)$$

In equation (3), f_{ab} is the relative frequency with which the symbols a and b have actually matched at the same position of a given set of pairwise alignments, while $f_a * f_b$ is the product of the relative frequencies of a and b in the same data set and therefore an estimation of the probability of seeing a and b aligned throughout all the alignments of the data set. If we consider f_{ab} to be an estimate of the probability of finding a and b matched in the data set, then it becomes possible to estimate the ratio of those two probabilities (their odds) and evaluate the extent to which a given substitution (match) between two symbols is over- or underrepresented in the alignments. The most notable property of log odds is to yield negative scores for events observed less often than expected by chance. In the context of optimal matching, this amounts to having a cost matrix that penalizes unexpected matches with negative values while expected matches or identities are rewarded with positive values. As in an alignment, two identical symbols do not systematically match, and the Dayhoff cost for substituting two identical symbols is often different from zero. In biology matching, various pairs of identical symbols can be associated with different positive values.

The rationale is that in biology, all conservations are not equally important. In the social sciences, however, the decision was made early to set conservation costs to 0 and substitution to variable costs. This model suggests that all social statuses are equally conserved, regardless of their nature. This may or may not be true. For instance, one may ask if the social cost should be the same for matching years as unemployed or years spent on the labor market. The equality of these statuses cannot be ruled out as long as it has not been formally demonstrated. For the time being, the proposed algorithm sticks to the mainstream procedure in the social sciences, but it would be trivial to adapt it so that different costs may be used for different types of identities.

To get a cost of zero for the substitution of two identical symbols, we use a normalized cost (*N_cost*) that is derived from the cost defined in equation (3) as follows:

$$N\_cost(a,b) = Dayhoff\ Cost(a,b) - \frac{Dayhoff\ Cost(a,a) + Dayhoff\ Cost(b,b)}{2}. \quad (4)$$

In equation (4), *Dayhoff cost* refers to the original Dayhoff cost (equation [3]) that is positive and maximized for identities while yielding lower (often negative) values for mismatches. A substitution matrix based on

N_costs has the same properties as the Dayhoff cost matrix except that it yields a null cost to the alignment of two identical sequences, a convenient property for cluster analysis based on a distance matrix. In biology, it is common practice to use log-odds matrices as a scoring scheme when applying the optimal matching algorithm. The main reason is that the versatility of the log-odds method makes it possible to discriminate between different types of mismatches in an objective and quantitative fashion.

As substitution and INDEL operations are mutually dependent, using cost matrices as defined in equation (3) or (4) calls for setting the value of the INDELs according to the cost matrix at hand, as shown in equation (5).

$$INDEL = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} C_{ij}}{(N^2 - N) \times 0.5}. \quad (5)$$

In equation (5), the cost for not matching a symbol (INDEL) was estimated using the Thompson formula (Thompson, Higgins, and Gibson 1994), where INDEL is set to the average substitution cost of the substitution matrix (i.e., the matrix average ignoring the values in the main diagonals). It is possible to distinguish two kinds of indels, the internal ones that occur between two given symbols and the terminal ones that come at the end of the shorter sequence to make its length equal to the longer one. In the context of this work, we simply attributed the INDEL value of equation (5) to internal indels only and lowered it to INDEL/2 for terminal ones, thus making it easier for indels to be terminal rather than internal.

Given a collection of sequences, the main difficulty is the proper estimation of an appropriate cost matrix. Using reference alignments is possible but may require some arbitrary knowledge. In the case of Dayhoff, using reference alignments was possible because closely related sequences were available whose alignment could be assembled in an unambiguous manner (i.e., without INDEL). In the social sciences, reference alignments are not available, and a strategy must therefore be worked out to generate them in a systematic and unbiased fashion. Over the last 15 years, several techniques have been introduced in biology and aimed at training position-specific substitution matrices through iterative sequence-alignment procedures (Lawrence et al. 1993; Hughey and Krogh 1996; Altschul et al. 1997; Bateman et al. 1999). PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool), one of the most popular tools in biology, is the one whose principle resembles most the one developed here. In PSI-BLAST, a biological sequence is first compared to all the others in the database, using an off-the-shelf substitution matrix. The

Figure 2
Pseudocode Describing the Iterative Training Procedure

```
0 - set CURRENT matrix to the identity matrix
I - Optimal matching of the sequences with the CURRENT matrix
II - Estimation of the NEW matrix on the alignments produced in (I)
Measure the Percent Identity on the alignments
Select alignments Yielding more than 60% identity
Count the matches/mismatches on the selected alignments
Weighting the counts of each alignment with its percent identity
Compute the NEW matrix
III–Comparison of NEW matrix and CURRENT matrix
If CURRENT==NEW, terminate
Else set CURRENT to NEW and proceed to I.
```

best alignments thus generated are selected according to their percentage of identity and used to update the matrix in a process that goes on, cycle after cycle, until successive cycles fail to modify the matrix, in which case the algorithm is said to have reached convergence. The proposed strategy is directly adapted from this iterative method and is outlined in the pseudocode shown in Figure 2.

In this context, the matrix can be viewed as a model used for generating optimal matches of the sequences. In other words, a correct matrix must be able to generate alignments similar to those it was estimated from. This equivalence is sought in the iteration procedure, in which matrices and alignments are alternatively generated until they both become invariant, suggesting an equivalent information content. Overall, this amounts to generating matrices whose purpose is to optimally summarize the information contained in the sequences. In this context, the alignments and the matrix can be viewed as two alternative models of the relationships among sequences. The convergence is meant to ensure that these two models are equivalent.

Empirical Cost Matrix Estimation of Social Science Data

Given a set of sequences of social statuses and a preestimated matrix (in this case, an identity matrix), pairwise alignments are generated with the *OMA* algorithm. This can be done either by exhaustively considering

all possible pairs of sequences or by restricting the training procedure (cf. Figure 2) to a random subset of the sequences if computation time is an issue. When computing matrix statistics from these alignments, the main caveat is the uneven alignment quality. While mismatches measured on almost identical strings can be expected to be meaningful substitutions, matches and mismatches measured on poorly matched strings may be suspicious. Dealing with low-quality alignments is a delicate issue in the social sciences as well as in biology. The simplest approach to deal with this limitation is to ignore alignments with a low percentage of identity, as done in PSI-BLAST (discussed previously). For instance, in the context of this article, we excluded all the alignments yielding less than 60 percent identity (equation [2]). Such a conservative threshold ensures the quality of the considered alignments and therefore the relevance of the observed substitutions.

Furthermore, based on strategies developed in biology, we also applied an extra weight on the selected alignments in order to ensure that the best alignments contribute more to the final matrix. This extra step is similar to the selection made for empirically estimating the costs of substitution (equation [2]), but it specifically helps smooth the convergence of the iterative process and also guarantees a stronger contribution of the most reliable alignments. We again used percentage of identity (as measured on the alignment) as a weighting scheme. This parameter is often regarded as a good indicator of correctness and was successfully used by Notredame, Holm, and Higgins (1998) to design local scoring schemes. We therefore used equation (6) to derive a collection of weights that are specifically applied to the relative frequency of each possible substitution associated with a given alphabet:

$$f_{ab} = \frac{\sum_1^S W_i * N_i(a, b)}{\sum_1^S W_i * L_i}. \tag{6}$$

Thus, weighting the relative frequency f_{ab} of symbol a matching symbol b in the alignments is estimated in the following fashion, where W_i is the weight associated with the alignment i, L_i is the length of that alignment, and $N_i(a, b)$ is the number of pairs ab perfectly matched in that alignment. The term $N_i(a, b)$ indifferently represents identical matches if a and b are the same symbols or mismatches if a and b are different symbols.

In equation (6), the weight W_i is meant to increase the contribution of trustworthy alignments, thus speeding the convergence process and decreasing the amount of noise contributed by spurious alignments. In the case of social science data in which sequence patterns are shaped

following less stringent rules than in biological sequences and therefore show more diversity, this approach allows us at the same time to take a greater variability of sequences into account and to limit the influence of outliers. To prevent possible underflow (i.e., division by 0) caused by a rare mismatch or match, a small value (0.001) is added to every frequency. Given frequencies tabulated for every possible pair of symbols, the substitution matrix is then computed using equation (3). This matrix is the new matrix that will be used in the next training round (cf. Figure 2).

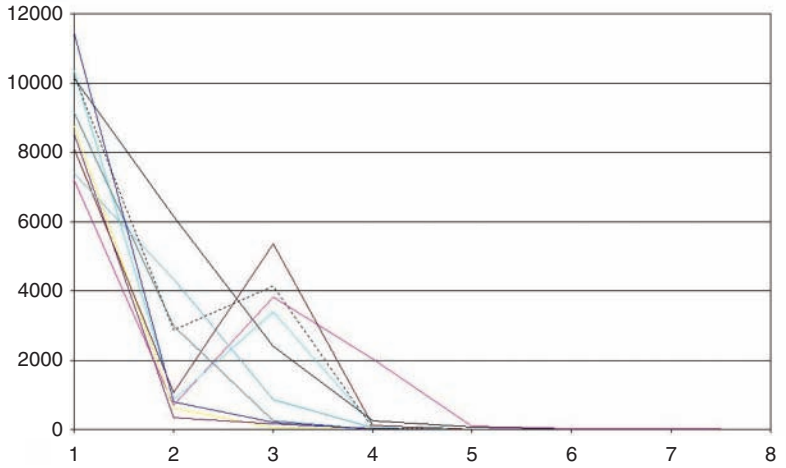
The iteration procedure is meant to optimize the cost matrix so that it summarizes as accurately as possible the information contained in the alignments from which it is estimated. That procedure is complete when a matrix is able to generate alignments with statistical properties similar to those it originates from. That convergence can easily be measured by estimating the difference between two successive matrices in the evaluation procedure (Δ), using, for instance, the mean square differences between them:

$$\Delta = \sqrt{\left(Avg\left((M_1(a,b) - M_2(a,b))^2\right)\right)}. \quad (7)$$

The iterative procedure is stopped when Δ becomes equal to 0. However, this procedure is merely an attempt to reach optimality, with no proven guarantee (Hughey and Krogh 1996). In this context, the simplest criterion to ensure optimality is to check that alternative trainings converge on similar matrices as indicated by low Δ values, as in equation (7). To validate this, we randomly selected 10 sets of 100 sequences in the test data set and trained the corresponding matrices, keeping the intermediate matrices obtained at every cycle. Figure 3 shows the average Δ measured between all these matrices against the iteration number.

Given a data set of 100 sequences 40 symbols long, Figure 3 shows the typical profile of several training procedures. The Δ is an estimation of the difference between the matrices of two successive rounds (low Δs indicate highly similar matrices). While Δ values tend to decrease over cycles, increasing values (peaks) usually result from the exceeding of a local minimum by the training procedure. Each curve in Figure 3 corresponds to one matrix estimation run. For each run, a set of 1,000 sequence pairs was chosen randomly (out of the 100*100 possible pairs) and kept through all the iterations. The results suggest that the estimation procedure is insensitive to this initial choice with a convergence systematically occurring in 5 to 6 cycles and final matrices highly correlated.[5] Altogether, this high correlation and the constant number of cycles suggest an efficient and robust training procedure.

Figure 3
Value of Δ Against the Number of Iterations of the Training Procedure of Substitution Costs for 10 Randomly Selected Sets of Sequences From the Swiss Household Panel Data



All the procedures described here have been encapsulated in a sequence analysis package called SALTT (Search Algorithm for Life Trajectories and Transitions). It can be compiled and installed on any UNIX-like platform including Linux, Cygwin, and Mac OSX. The package and its documentation are distributed under the General Public License and available free of charge from the authors (Notredame et al. 2005).

Criteria for Comparing Outcomes From Various Cost-Setting Strategies

To test whether the proposed solution provides more adequate results than previous methods, one may consider some criteria specific to the training procedure as well as a set of criteria widely used to estimate how well sequence analysis and cluster analysis perform.

The first and simplest criterion to establish the validity of the proposed strategy is to apply it to biological sequences and train matrices that could be compared to standard biological matrices. We have done so on a

Figure 4
Random Splitting of Symbol A into Two New Symbols (M and N) Not Belonging to the Original Alphabet

```
Seq1:  AABDEEBADBEDA Seq2:  BDAEAEAADBADA

A -> M ou N
Seq1b: MNBDEEBNDBEDM Seq2b: BDNEMEMNDBNDM
```

well-known collection of 500 related human sequences known as the kinome (Manning et al. 2002). The procedure delivered a substitution matrix highly correlated to a standard point accepted mutation in which all the known mutational preferences between amino acids could easily be recognized. We then do the same by comparing three distinct sets of social sciences data representing the same sequential reality.

Then, the training procedure is evaluated by testing its ability to correctly identify the closeness of two different symbols, using solely the information contained in the data. To do so, we use a set of sequences to compare a reference cost of substitution between two given symbols produced by the training procedure (e.g., AB), with the cost produced by the training procedure for the same substitution, in the case where one of the symbols (e.g., A) has been randomly split into two new symbols (M and N) not belonging to the alphabet. As symbols M and N are actually "hidden A," we expect the training procedure to determine the substitution costs AB, MB, and NB as equivalent. Figure 4 shows for two given sequences how a symbol is randomly split into two new symbols not belonging to the original alphabet.

Testing the Quality of the Clustering

A third set of criteria pertains to quality testing of cluster analysis. One of the main difficulties with clustering methods lies in the determination of the number of clusters really present in the data (Milligan and Cooper 1985, 1987). There is no perfect method to establish this number, but several indicators may be used to help decide (Everitt 1979; Bock 1985; Hartigan 1985; Milligan and Cooper 1985; SAS Institute 2004). For

Milligan and Cooper (1987), there are two categories of tests concerning the quality of cluster analysis: The first considers that internal criteria are able to validate the results of the clustering, that is, to justify the number of clusters chosen. The second one uses external criteria. Such criteria represent information that is external to the cluster analysis and was not used at any other point in the cluster analysis (Milligan and Cooper 1987).

In terms of internal criteria, Milligan and Cooper (1985) have evaluated and compared 30 statistics known as *stopping rules* that help in deciding how many "real" clusters are present in the data. The availability of such indices in main statistical software packages (such as SAS or SPSS) is of course a nonnegligible element of choice concerning what criteria to use. Two of the most efficient indices among the 30 that Milligan and Cooper (1985, 1987) have evaluated are part of the SAS software. The first one is a pseudo-F developed by Calinski and Harabasz (1974); it represents an approximation of the ratio between intercluster and intracluster variance. The second index is expressed as $Je(2)/Je(1)$ (Duda and Hart 1973) and may be transformed into a pseudo-t^2.[6] The third criteria we used is R^2, which expresses the size of the experimental effect. It is reasonable to look for a consensus among the three criteria (Nargundkar and Olzer 1998; SAS Institute 2004). We can then define the stopping rule for a statistically optimal cluster solution as a local peak of the pseudo-F (high ratio between inter- and intracluster variance), associated with a low value of pseudo-t^2 that increases at the next fusion and a marked drop of the overall R^2.[7] Generally, a cluster solution is said to be statistically optimal when the number of classes is kept constant across strategies, when the intercluster variance is highest, and when the intracluster variance is lowest. Put another way, clusters should exhibit two properties, external isolation and internal cohesion (Punj and Stewart 1983). Therefore, using comparative scree plots is a straightforward way of dealing with the issue of testing cluster solutions drawn from distances based on various cost schemes, including the computationally derived one. A given cluster solution is retained for analysis only if at least two among those three criteria (pseudo-F, pseudo-t^2, and R^2) support its validity.

External criteria refer to the extent to which clusters drawn from the data correlate with either independent variables or outcomes (Milligan and Cooper 1987). Clusters that do not associate with these variables are of little help in social research as the ultimate goal of social sciences is explanation rather than description. A third criterion is more intuitive: To what extent are empirical clusters easily comprehended, based on prior knowledge of the phenomenon and the central hypothesis of the research? This criterion

can be approached by using experts and computing interreliability estimates. The procedure in that case is as follows: Provide cluster solutions based on the various cost schemes, and have a set of raters who decide independently which is their favorite solution. Then one may compute interrater reliability and see which coding scheme comes up first in the list.

Given the importance of the debate concerning the influence of sociostructural factors on the occupational trajectories of women in the sociological field and the availability of high-quality data on occupational status during entire life courses, we test these methods on data sets addressing this topic.

Description of the Test Samples

Considering the fact that women's labor market participation is more diverse than that of men (Myrdal and Klein 1956; Levy 1977; Mott 1978; Elder 1985; Moen 1985; Höpflinger, Charles, and Debrunner 1991; Moen and Yu 2000; Blossfeld and Drobnic 2001; Krüger and Levy 2001; Levy, Widmer, and Kellerhals 2002; Moen 2003; Widmer, Kellerhals, and Levy 2003; Bird and Krüger 2005; Levy, Gauthier, and Widmer 2006), and in order to facilitate the comparisons between the data sets, for each database we selected only women who were married or living with a partner at the time of the interview. Moreover, in order to maximize the quality of the data, we retain only the trajectories that had less than 10 percent of missing values.

Sample Test 1: Social Stratification, Cohesion, and Conflict in Contemporary Families

The first sample of occupational trajectories is drawn from a retrospective questionnaire of the study Social Stratification, Cohesion, and Conflict in Contemporary Families (SCF) that was conducted in 1998 with 1,400 individuals living as couples in Switzerland (Widmer, Kellerhals, and Levy et al. 2003; Widmer, Kellerhals, and Levy 2004). Respondents were asked to provide information about every year of their occupational trajectory starting from age 16, onward to 64. Every year of the trajectories was coded using a seven-category code scheme: full-time employment, part-time employment, positive interruption (sabbatical, trip abroad, etc.), negative interruption (unemployment, illnesses, etc.), housework, retirement, and full-time education. Data were right truncated as most individuals had not yet reached the age of 64 at the time of the interview. Sociostructural indicators (socioeconomic status of orientation family, educational

level, number of children, and income) were measured for the time of the interviews only. The final sample size was 564 women.

Sample Test 2: The Swiss Household Panel

Since 1999, the Swiss Household Panel (SHP) has collected data on a representative sample of private households in Switzerland on a yearly basis.[8] In its third wave, the SHP included a retrospective questionnaire sent to 4,217 households (representing 8,913 individuals). For reasons of validity, the analysis of the subsample of individuals who answered the retrospective questionnaire was restricted to those aged 30 and older, decreasing the sampled female population to 1,935. The SHP asked respondents to provide information on their educational and occupational status from their birth to the present. Each change in status is associated with a starting year and an ending year. We recoded these the same way as for Sample Test 1. Sociostructural indicators comparable to those in Sample 1 were also obtained. This sample included 1,107 women.

Sample Test 3: Female Job Histories From the Wisconsin Longitudinal Study

The Wisconsin Longitudinal Study (WLS) is a long-term study of a random sample of 10,317 men and women who graduated from Wisconsin high schools in 1957. This data set is for public use and available at the University of Wisconsin–Madison Web site (http://www.ssc.wisc.edu/wlsresearch). The female job histories of 1957-1992 were constructed by Sheridan (1997) from the 1957, 1964, 1975, and 1992 WLS data collections. The data also include social background, youthful aspirations, schooling, military service, family formation, labor market experiences, and social participation. The "female job histories" data concern 5,042 women born in 1938 and 1939. We could retain only three main occupational statuses, namely, full-time paid work, part-time paid work, and full-time housewife. There were 2,243 women in this sample.

Results

Production of Data-Driven Costs of Substitution

From a sociological point of view, we could expect a relative stability of the costs of substitution from one set of sequences to another, the

occupational trajectories of contemporary Swiss and North American women being to a certain extent comparable, at least in regard to the influence of the birth of children on the reduction or cessation of paid work. The individual sequences of occupational statuses are built by attributing a single symbol (a code corresponding to a given occupational status) to each year of life of the respondents.[9] Table 1 compares the different costs of substitution either set arbitrarily to identity, following theoretical arguments concerning differences among types and rates of occupational activities (for details, see Widmer, Levy, et al. 2003), or by means of a training procedure in the different databases.

Table 1 shows that the training procedure produces costs that are more differentiated than identity costs. The range of costs is also broader, partly because the procedure is sensitive to very rare substitutions. The stability of the trained costs of substitution from one database to another confirms the ability of the training to produce meaningful cost schemes. The training procedure reflects some relations between the different statuses that are sociologically relevant. Compared to identity costs that may not be differentiated between men and women, the trained costs reveal, for example, the relative ease (the low costs) with which women in the samples go from paid work to housework. The comparison of knowledge-based costs and trained costs of substitution shows a high similarity between the two sets of values, which are correlated at .68 ($p < .01$) with trained costs for SCF data, at .63 ($p < .01$) for SHP data, and at .73 ($p < .05$) for WLS data. Table 2 shows Pearson's coefficient of correlation between the costs by method of cost setting and database.

Table 2 shows that the trained costs of substitution are more strongly associated with each other from one data set to another than they are with costs set to either identity or to knowledge-based values. On the other hand, even if it remains relatively high, the associations between trained, knowledge-based, and identity costs are systematically weaker than those between trained costs. This confirms the stability of the results stemming from the training procedure and explains at least partly the slightly but systematically different (and more highly correlated) results it provides compared to the two other strategies (identity and knowledge based).

Validation of the Training Procedure

An important issue in the use of a computerized data–based determination of substitution costs is to assess the extent to which this procedure is able to process information in a sociologically relevant way. Three

Table 1
Comparisons of Identity, Knowledge-Based, and Trained Costs of Substitution for Three Data Sets: SCF, SHP, and WLS

	Costs of Substitution				
Substitutions of Occupational Status	Identity	Knowledge Based	Trained SCF	Trained SHP	Trained WLS
Full-Time * Part-Time	1.0	0.8	0.6	0.5	0.5
Full-Time * Negative Interruption	1.0	1.0	0.7	0.7	0.4
Full-Time * Positive Interruption	1.0	0.8	0.7	0.7	
Full-Time * At Home	1.0	1.0	0.5	0.5	
Full-Time * Retirement	1.0	1.0	0.9	0.8	
Full-Time * Education	1.0	1.0	0.5	0.5	
Full-Time * Missing	0.3	0.3	0.6	0.6	0.5
Part-Time * Negative Interruption	1.0	0.8	0.7	0.7	0.5
Part-Time * Positive Interruption	1.0	0.8	0.9	0.8	
Part-Time * At Home	1.0	1.0	0.5	0.5	
Part-Time * Retirement	1.0	0.8	1.0	0.8	
Part-Time * Education	1.0	0.8	0.7	0.7	
Part-Time * Missing	0.3	0.3	0.7	0.7	0.5
Negative Interruption * Positive Interruption	1.0	1.0	1.5	1.2	
Negative Interruption * At Home	1.0	1.0	0.7	0.8	
Negative Interruption * Retirement	1.0	1.0	1.3	0.9	
Negative Interruption * Education	1.0	0.8	0.8	0.9	
Negative Interruption * Missing	0.3	0.3	0.9	1.0	0.5
Positive Interruption * At Home	1.0	1.0	0.9	0.8	
Positive Interruption * Retirement	1.0	1.0	1.1	1.4	
Positive Interruption * Education	1.0	0.8	0.8	0.8	
Positive Interruption * Missing	0.3	0.3	0.9	0.9	
At Home * Retirement	1.0	1.0	1.0	0.8	
At Home * Education	1.0	1.0	0.6	0.7	
At Home * Missing	0.3	0.3	0.6	0.7	
Retirement * Education	1.0	1.0	1.5	1.6	
Retirement * Missing	0.3	0.3	1.3	1.5	
Education * Missing	0.3	0.3	0.7	0.7	
Insertion or Deletion	0.5	0.5	0.8	0.8	0.5

Note: SCF = Social Stratification, Cohesion, and Conflict in Contemporary Families; SHP = Swiss Household Panel; WLS = Wisconsin Longitudinal Study.

different tests were used. The first one referred to the ability of the procedure to evaluate the closeness of a symbol belonging to the alphabet with an unknown symbol not belonging to it. The second one focused on the degree of agreement between classifications of social trajectories made by

Table 2
Pearson's Correlation Between Costs Matrices, by Method of Cost Setting and (Full) Data Sets

	Identity	Knowledge	SCF Trained	SHP Trained	WLS Trained
Identity	1.00	.98***	.66***	.61***	.71*
Knowledge based	.98***	1.00	.68***	.63***	.73*
SCF trained	.66***	.68***	1.00	.96***	.97***
SHP trained	.61***	.63***	.96***	1.00	.94***
WLS trained	.71*	.73*	.97***	.94***	1.00

Note: UNIX command line to produce the trained matrix: saltt -e '-in dataset.dat -action + pavie_seq2pavie_mat _TGEPF50_THR60_TWE04_SAMPLE50000_'. SCF = Social Stratification, Cohesion, and Conflict in Contemporary Families; SHP = Swiss Household Panel; WLS = Wisconsin Longitudinal Study.
*$p < .05$. ***$p < .001$.

specialists in the field compared with classifications of the same data based on identity, knowledge-based, and trained costs of substitution. The third one consisted of measuring the extent to which clusters drawn from the data correlate with some independent sociostructural variables or outcomes.

Identifying the Proximity of Unknown Symbols

A first way of validating the training procedure consists of measuring the extent to which it is able to unravel the proximity of two given symbols, based on no other information than the data itself. We tested this for the SCF set of sequences by randomly replacing a given symbol of the sequences alphabet A = {A, B, C, D, E, F, G, X}, which corresponds in this case to an occupational status, with two symbols that did not belong to the original alphabet of that set of sequences, that is, symbols whose actual identity was hidden.

Using the training procedure, we then compared the original costs for substituting, for example, Symbol A with Symbol B, with the costs we obtained after having randomly replaced every A with either the hidden symbol M or N (cf. Figure 4). In a second run, we did the same by replacing each B with the hidden symbol O or P. We finally got five different expressions of the same initial substitution (in this example, AB = NB = MB = AO = AP), each associated with a specific cost. This procedure was applied to all pairs of symbols of the data set in turn. If we consider E_i and E_j to be respectively the

*i*th and *j*th elements of the original alphabet and their two random substitutes—respectively S1(*Ei*), S2(*Ei*) and S1(*Ej*), S2(*Ej*)—there are five costs of substitution to take into account if we consider only the substitutions involving at least one symbol belonging to the original alphabet. Under these conditions, as they are actually different expressions of the same initial substitution, we should expect those five trained costs to be identical, or at least close to each other. To compare all those values in a synthetic way for the entire alphabet, we computed a standardized difference between the trained costs of substitution associated with a given pair of symbols belonging to the original alphabet and the trained costs of substitution between one of those original symbols and the substitute of the other one, as shown in equation (8).

$$\textit{Std Difference} = \frac{(\text{cost}(\text{E}_i[\text{S}_1(E_j)]) - \text{cost}(\text{E}_i\text{E}_j) + (\text{cost}[\text{S}_1(E_i)]\text{E}_j) - \text{cost}(\text{E}_i\text{E}_j))}{2*\text{cost}(\text{E}_i\text{E}_j)}. \quad (8)$$

The proximity of the five substitution costs associated with a given original pair of symbols and their substitutes was compared in two ways, using either the first substitute of that pair of symbols (as shown in equation [8]) or the second one (where S_2 replaces S_1 in equation [8]). All those values were then tabulated in Table 3. Its lower part contains the standard differences between the substitutions of E_i, E_j, and their first substitute (cost E_iE_j compared to $E_i[S_1(E_j)]$ and $[S_1(E_i)]E_j$), whereas the upper part contains the values associated with their second substitute (cost E_iE_j compared to $E_i[S_2(E_j)]$ and $[S_2(E_i)]E_j$).

Table 3 shows clearly that the training procedure identifies very precisely the closeness of two distinct, but actually identical, symbols.[10] Among the 56 different costs of substitution in Table 3, 49 (87 percent) show a difference not larger than 10 percent compared with the original cost. The greater differences may be attributed to the fact that the training procedure is relatively sensitive to rare symbols. For example, symbols C, D, F, and X represent altogether only about 2 percent of the total symbols used in the sequences. The great majority of the hidden costs differing notably from their original costs are concerned with such rare symbols. The difference is maximal when it concerns two rare symbols. The ability of the training procedure to identify the similarity of two unknown symbols based on the data set at hand is one of the main strengths of this way of setting costs of substitution. Even if it stays relatively close to identity costs of substitution, this procedure takes into account the real relations of the different symbols present in the sequences and is therefore highly informative. On one hand, it avoids particular relationships remaining undetermined; on the other hand, it works as a predictive tool in the sense

Table 3
Standardized Difference Between the Original Trained Costs of Substitution and Their Substitutes

	A (%)	B (%)	C (%)	D (%)	E (%)	F (%)	G (%)	X (%)	Relative Frequency (%)
A	—	8	7	6	0	0	10	0	33.5
B	0	—	0	6	0	25	7	7	19.5
C	7	0	—	0	–7	9	0	–15	0.5
D	6	6	5	—	0	11	0	–28	1.0
E	0	–10	–7	–6	—	13	8	0	31.0
F	6	10	0	0	13	—	0	0	0.1
G	10	0	0	6	8	0	—	–7	14.0
X	0	0	–15	6	0	0	–7	—	0.4

Note: Rows and columns are given the name of a symbol belonging to the alphabet, although each cell of the table compares the substitution cost of three pairs of symbols (the original one and two substitutes) according to equation (8).

that two different symbols with low substitution costs can be predicted to substitute easily for one another in real life.

Automatic Versus Classification by Judges

Another way to validate the training procedure is to test the extent to which automatic classification succeeds in replicating a classification of sequences made by experts on a small subset of well-identified sequences. To do so, we extracted a sample of 100 occupational trajectories of women from each data set. Four judges were asked to classify them in a number of clusters that corresponded to previous empirical findings (Widmer, Levy, et al. 2003; Levy, Gauthier, and Widmer 2006) and to theoretical schemes (i.e., Kohli 1986). In each case, we retain only the sequences that were classified the same way by at least three (out of four) judges. The interrater agreement lies between 83 percent and 88 percent.

To keep the computation procedures as parsimonious as possible, we first exactly replicated with SALTT the results we obtained with TDA using two different cost settings (identity and knowledge based). That allowed us to produce optimal alignments and to compare the distance matrices for the three strategies (identity, knowledge based, and training) from within SALTT.

For each set of sequences, we ran three optimal matching analyses, the first one using identity costs of substitution (for details, see above), the

Table 4
Association (khi^2 and Symmetric) Between Judges and Automatic Classification, by Method of Cost Setting

	Database		
Method	SCF	SHP	WLS
Identity * Judges	213.2454*	213.4108*	143.9678*
	(khi^2, df = 16)	(khi^2, df = 16)	(khi^2, df = 9)
	Λ symmetric	Λ symmetric	Λ symmetric
	0.8034 (value)	0.7500 (value)	0.7037 (value)
	0.0458 (ASE)	0.0582 (ASE)	0.0684 (ASE)
Knowledge Based * Judges	206.1951*	228.4631*	148.6864*
	(khi^2, df = 16)	(khi^2, df = 16)	(khi^2, df = 9)
	Λ symmetric	Λ symmetric	Λ symmetric
	0.8120 (value)	0.7705 (value)	0.7196 (value)
	0.0440 (ASE)	0.0623 (ASE)	0.0677 (ASE)
Trained * Judges	224.5436*	235.1387*	143.2652*
	(khi^2, df = 16)	(khi^2, df = 16)	(khi^2, df = 9)
	Λ symmetric	Λ symmetric	Λ symmetric
	0.8291 (value)	0.7797 (value)	0.7037 (value)
	0.0434 (ASE)	0.0602 (ASE)	0.0677 (ASE)

Note: ASE = Asymptotic standard error; SCF = Social Stratification, Cohesion, and Conflict in Contemporary Families; SHP = Swiss Household Panel; WLS = Wisconsin Longitudinal Study.
*$p < .001$

second one using knowledge-based costs, and the third one using costs stemming from the training procedure. A distance matrix was computed for each set of sequences and for each strategy and then entered into a cluster analysis. Table 4 shows the degree of association of khi^2 and Λ (Goodman and Kruskal 1979; Olszak and Ritschard 1995) between the clusters made by the judges and those stemming from automatic classification.

Table 4 shows that results provided by a trained matrix lead to significant associations with the classification by judges for the three data sets considered. For the Wisconsin study, results are about the same when using either identity or trained costs of substitution. Trained costs never lead to a weaker association (Λ symmetric) with judges' classifications than identity costs or knowledge-based costs for the SCF and SHP data sets. Results are less straightforward concerning the WLS data, with knowledge-based costs performing slightly better than trained costs. The fact that Wisconsin data are less differentiated (sequences with only three

different statuses as opposed to seven in the other databases and respondents all about the same age) may explain why trained costs do not lead to a markedly different solution than the two alternative strategies. In all cases, the associations are quite high and significant, suggesting the ability of the method to provide meaningful cost schemes. Given the fact that the reference classification based on judges responses was very consensual and based on predefined categories, results of that test express the ability of the procedure to differentiate clear-cut, sociologically relevant categories out of the data rather than to evaluate the extent to which those results and the underlying costs of substitution reflect the theoretical and subjective conceptual frame of an expert.

Association With External Criteria

A third validation procedure consisted of measuring the extent to which clusters drawn from the data correlate with either independent sociostructural variables or outcomes (Milligan and Cooper 1987; Rapkin and Luke 1993), an approach that seemingly few studies have used so far (Milligan and Cooper 1987). Clusters that do not associate with these variables are of little help as the ultimate goal of social sciences is explanation rather than description.

For each strategy, the three stopping-rule criteria aimed at determining the number of clusters in the data (pseudo-t^2, pseudo-F, and R^2) suggested the presence of three clusters in the SCF and SHP data and of four clusters in the WLS data. A closer look at the data reveal that those clusters correspond precisely to typical female trajectories, as described elsewhere (Moen 1985; Höpflinger et al. 1991; Erzberger and Prein 1997; Widmer, Levy, et al. 2003; Levy et al. 2006), namely, trajectories characterized by full-time employment, part-time employment, and full-time as a housewife. In the Wisconsin data, the clusters are the same, but with a fourth one representing a return to the labor market after a period at home. Such a cluster also appears when the clusters of SCF and SHP data are further subdivided. The greater homogeneity of WLS data in terms of age of respondents and completeness of the sequences (no right truncatures) may explain the better visibility (consensus between stopping-rule criteria) of that fourth cluster, which is also documented in the literature (Widmer, Levy, et al. 2003; Levy et al. 2006).

We first ran a multinomial logistic regression[11] on each data set (SFC, SHP and WLS), using cluster membership (which represents in this case types of occupational trajectories) as response variables and a set of

Table 5
khi^2 Values of the Likelihood Ratio Test by Database and Cost-Setting Method

		Cost-Setting Method					
		Identity		Knowledge Based		Trained	
Data Sets	*df*	khi^2	$p >$ khi^2	khi^2	$p >$ khi^2	khi^2	$p >$ khi^2
Set 1: SCF data, 3 clusters	272	290.19	.2143	280.87	.3428	288.60	.2339
Set 1: SCF data, 5 clusters	596	553.02	.8956	547.03	.9250	522.11	.9867
Set 2: SHP data, 3 clusters	404	568.36	< .0001	562.04	< .0001	512.34	< .0002
Set 2: SHP data, 5 clusters	808	897.67	.0150	863.32	.0865	740.12	.9574
Set 3: WLS data, 4 clusters	258	307.35	.0189	323.81	.0034	288.37	.0939

Note: SCF = Social Stratification, Cohesion, and Conflict in Contemporary Families; SHP = Swiss Household Panel; WLS = Wisconsin Longitudinal Study.

indicators of social positioning (socioeconomic position of the orientation family, including level of education, number of children, age, and household income) generally considered (cf. description of the sample) as intervening variables in shaping female occupational trajectories. To be consistent with the stopping-rule criteria—that is, a consensus between pseudo-t^2, pseudo-F, and R^2—we retained in this first step the three-cluster solutions that those criteria pointed out for each data set. As they are more homogeneous, they represent about the same social reality in each data set and therefore remain sociologically relevant. We then performed the tests on the five-cluster solutions for SCF and SHP data to check the efficiency of the different cost-setting methods on other empirically founded classifications (Widmer, Levy, et al. 2003; Levy et al. 2006). We felt justified in doing this because two new clusters emerged from further subdivision of the first three clusters defined by the proposed criteria (R^2, pseudo-F, and pseudo-t^2). Table 5 shows the test of likelihood ratio applied to those multinomial regressions. The likelihood tests compare a given model with the saturated one (a model that exactly replicates the data), meaning in this case that the smaller the value of khi^2 (i.e., the larger the p value), the better the model fit to the data.[12]

One can read from Table 5 that the trained costs of substitution allow building a model that fit better to the data in all cases compared to identity costs and in four out of five cases compared to knowledge-based costs. Put another way, clusters produced by trained costs of substitution are

more sensitive to predictors than clusters produced by either identity costs or knowledge-based costs. This is true, although not with the same strength, for the three sets of sequences. The fit is significantly better (i.e., the model stemming from trained costs does not differ significantly from the saturated model, whereas the two others do) in two cases and with two data samples.

Discussion

Setting costs of substitution in the process of aligning sequences of social statuses is controversial because it may significantly influence the results of the analysis. We propose a method to determine costs of substitution empirically, which we tested using three distinct sets of social science data. The training procedure that we present appears to be, to our knowledge, the only one that is exclusively empirically grounded and optimized.

First, we considered the correlation between the substitution matrices for a given alphabet across three data sets of the social sciences realm representing the same social realities (sequences of occupational statuses along the life course) and three cost-setting strategies. The training procedure leads to results that are very similar to those stemming from the two other methods (substitution costs represented as an identity matrix or following some knowledge-based weighting). In this sense, cost variability did not appear to modify the general results of the analysis. Nevertheless, the costs stemming from the training procedure may claim a greater legitimacy as they reflect the actual relationships of the symbols considered. That legitimacy is reinforced by the very high correlation existing between the substitution matrices stemming from the training procedures applied to the three data sets at hand. In this sense, the values of the trained cost matrices may even be considered as a validation a posteriori of the use of alternative costs of substitution (knowledge based or identity) found in the literature for the specific case of occupational trajectories. Moreover, the training procedure shows some interesting features that should be further explored, such as the possibility to differentiate specific substitution costs according, for example, to gender. The ability of the trained costs to provide a clustering that is better associated to some sociologically unequivocal classification of reference than the identity costs and the knowledge-based costs did illustrate the ability of the training procedure to discover some structural features of the data that are sociologically relevant.

Second, based on likelihood ratio tests of multinomial logistic regressions, we compared the associations between cluster solutions (response variables) and a set of relevant sociostructural variables (intervening variables) for the three cost-setting strategies across the three data sets at hand. Here again, the training procedure led to better results than the identity and the knowledge-based costs did. That is, the data-driven costs of substitution contributed to classifications that fit better with widely recognized sociological models of women's labor market participation than the two other strategies. Taking into account the actual structure of the data provides models that fit better with external factors than undifferentiated or knowledge-based costs schemes.

Finally, the ability of the training procedure to discover certain actual internal relationships of the data and therefore to offer an efficient and empirically grounded way to determine costs of substitution is demonstrated in another way as it is able to accurately identify the closeness of two formally identical, but artificially differentiated, substitution costs (here, between two occupational statuses). Moreover, the degree of closeness between the substitution costs is also informative about the relative proximity of the symbols and the sociological reality they represent.

The training procedure offers significant improvements compared to the methods generally used until now in social sciences. By revealing every symmetric relationship among those symbols, this procedure avoids assigning a cost based on prior knowledge that would later appear to be erroneous when compared to the actual data. The results show that for any pair of symbols of a given alphabet, the produced trained costs of substitution remain remarkably similar from one data set to another. It means that those costs do reflect some important information concerning the actual (in this case, social) significance of the symbols constituting the sequences and do not represent just abstract values varying from data set to data set (or from one training session to another). Therefore, these costs also constitute a predictive feature, in the sense that two different symbols with low substitution costs can be predicted to easily substitute for one another in real life. Identification of these low substitution costs therefore make it possible to predict situations likely to occur in similar contexts at similar ages and at similar frequencies.

In comparison with approaches based on transitions costs, which are computed within each single sequence taken separately, the proposed method aims to determine substitution costs by looking for a match or mismatch at each specific position throughout all pairs of sequences. In this sense, the latter method is based on richer information and grants a

higher importance to time (i.e., to age and social age) and to the relations between sequences than cost schemes based on transitions rates.

There is on one hand a constant and clear similarity between the results stemming from the three cost-setting strategies (identity, knowledge based, and training) and on the other hand a significant improvement in the tests of internal and external validity of the results provided by the training procedure. The conditions under which the method is most appropriate remain to be systematically tested. The experiments presented in this article point in several directions. First, the method provides a strong leverage when no or few theoretical arguments may be brought up to the scene in support of a cost solution or when contradicting theories propose different cost schemes. In other words, it is best suited for an exploratory research design. Second, this method is ideal whenever too many statuses have been used to characterize the data. We show, for instance, in this study that the proposed procedure reveals the identity between two statuses that may have been coded separately. Finally, the cost estimation provides a means for quantifying the relationship among symbols; as such, it can be used to identify and discover equivalences among categories. In itself, this means of quantification may prove to be a useful investigative tool for the social sciences.

There are several limitations to the solution proposed in this article. First, the method deals poorly with symbols occurring rarely in sequences. Whenever this happens, the estimations of substitution costs are less accurate and more variable. Second, a key property of the optimal matching algorithm is to rely on the assumption that events defining a life trajectory are chronologically ordered and collinear among the considered sequences. This is, of course, a simplification, but it seems to hold reasonably well when considering sequences with a high percentage of identity. However, it should be mentioned that if recurrent subsequences were to be found scattered in different periods of life, they could probably be recovered using techniques related to the one that we describe here, such as Gibbs sampling (Lawrence et al. 1993; Abbott and Barman 1997) or the local alignment algorithm (Smith and Waterman 1981). Second, this algorithm, like other optimal matching algorithms, assumes the independence of each position constituting a sequence. This may be oversimplifying as one can argue that life trajectories are not homogeneous. They may be substructured in smaller units (life stages, transitions, turning points, specific life events, etc.), whose sizes may vary but should be kept intact in the alignments. This issue is likely to arise when comparing very distinct sequences. When this situation occurs, it may be worthwhile to modify the

proposed algorithm. Nevertheless, the issue remains to automatically identify meaningful borders defining those subsequences. In biology, multiple sequence alignments have been used successfully to identify the exact extent of subsequences conserved across related sequences (Notredame, Higgins, and Heringa 2000). It is certainly worthwhile to explore the potential of this method in the social sciences.

Notes

1. This freeware is available from the Ruhr-Universität Bochum Web site at http://steinhaus.stat.ruhr-uni-bochum.de/tda.html.

2. This freeware is available from the University of Chicago Web site at http://home.uchicago.edu/aabbott/om.html.

3. This freeware is available from the Strasbourg Bioinformatics Platform Web site at ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX.

4. "Thus, while substitution must be carefully handled, it is not a supersensitive task whose errors will be compounded by later stages in the analysis" (Abbott and Hrycak 1990:176).

5. Student's t tests performed on the 10 values generated by the training procedures for each cost of substitution reveal that those values do not differ from the mean ($p < .0001$, $df = 9$).

6. Hotelling's T^2 is a statistical measure of the multivariate distance of each observation from the center of the data set. This is an analytical way to find the most extreme points in the data.

7. This is the ratio between interclass variance and total variance.

8. This data set is for public use. Access to the data is provided by the Swiss Household Panel (SHP) Web site at http://www.swisspanel.ch.

9. Following the availability of the data, the range considered is 16 to 65 years old for Social Stratification, Cohesion, and Conflict in Contemporary Families and SHP data, and 20 to 56 years old for Wisconsin Longitudinal Study data.

10. Spearman correlation coefficient $= .734$ ($p < .01$).

11. We used PROC CATMOD of the SAS software.

12. At $p \leq .05$, the tested model is not statistically different than the saturated one.

References

Abbott, Andrew. 1984. "Event Sequence and Event Duration: Collocation and Measurement." *Historical Methods* 17:192-204.

Abbott, Andrew. 1990a. "Conception of Time and Events in Social Science Methods: Causal and Narrative Approach." *Historical Methods* 23:140-50.

Abbott, Andrew. 1990b. "A Primer on Sequence Methods." *Organization Science* 1:375-92.

Abbott, Andrew. 1995a. "A Comment on 'Measuring the Agreement Between Sequences.'" *Sociological Methods & Research* 24:232-43.

Abbott, Andrew. 1995b. "Sequence Analysis: New Methods for Old Ideas." *Annual Review of Sociology* 21:93-113.

Abbott, Andrew. 2001. *Time Matters: On Theory and Method*. Chicago: University of Chicago Press.

Abbott, Andrew and Emily Barman. 1997. "Sequence Comparison Via Alignment and Gibbs Sampling: A Formal Analysis of the Emergence of the Modern Sociological Article." *Sociological Methodology* 27:47-87.

Abbott, Andrew and John Forrest. 1986. "Optimal Matching Methods for Historical Sequences." *Journal of Interdisciplinary History* XVI:471-94.

Abbott, Andrew and Alexandra Hrycak. 1990. "Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers." *American Journal of Sociology* 96:144-85.

Abbott, Andrew and Angela Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology." *Sociological Methods & Research* 29:3-33.

Aisenbrey, Silke. 2000. *Optimal Matching Analyse. Anwendungen in Den Sozialwissenschaften* (Optimal Matching Analysis: Applications in the Social Sciences). Opladen, Germany: Leske + Budrich.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jianzhi Zhang, Zhu Zhang, Webb Miller, and David Lipman. 1997. "Gapped Blast and Psi-Blast: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25:3389-3402.

Bateman, Alex, Evan Birney, Richard Durbin, Sean R. Eddy, Robert D. Finn, and Erik L. Sonnhammer. 1999. "Pfam 3.1: 1313 Multiple Alignments and Profile HMMs Match the Majority of Proteins." *Nucleic Acids Research* 27:260-62.

Billari, Francesco C. 2001. "Sequence Analysis in Demographic Research and Applications." *Canadian Studies in Population* 28:439-58.

Bird, Katherine and Helga Krüger. 2005. "The Secret of Transitions: The Interplay of Complexity and Reduction in Life Course Analysis." Pp. 173-94 in *Towards an Interdisciplinary Perspective on the Life Course*, vol. 10, edited by R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, and E. Widmer. Amsterdam: Elsevier JAI.

Blair-Loy, Mary. 1999. "Career Patterns of Executive Women in Finance: An Optimal Matching Analysis." *American Journal of Sociology* 104:1346-97.

Blossfeld, Hans-Peter and Sonja Drobnic. 2001. *Careers of Couples in Contemporary Society: From Male Breadwinner to Dual Earner Families*. New York: Oxford University Press.

Bock, Hans H. 1985. "On Some Significance Tests in Cluster Analysis." *Journal of Classification* 2:77-108.

Calinski, Tadeusz and Joachim Harabasz. 1974. "A Dendrite Method for Cluster Analysis." *Communication in Statistics* 3:1-27.

Chan, Tak Wing. 1995. "Optimal Matching Analysis: A Methodological Note on Studying Career Mobility." *Work and Occupations* 22:467-90.

Dayhoff, Margaret O., Robert M. Schwartz, and Bruce C. Orcutt. 1978. "A Model in Evolutionary Change in Proteins." Pp. 345-52 in *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, edited by M. O. Dayhoff. Washington, DC: National Biomedical Research Foundation.

Dijkstra, Wil and Toon Taris. 1995. "Measuring the Agreement Between Sequences." *Sociological Methods & Research* 24:214-31.

Duda, Richard O. and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis*. New York: John Wiley.

Durbin, Richard, Sean E. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press.

Elder, Glen H. 1985. *Life Course Dynamics: Trajectories and Transitions, 1968-1980*. Ithaca, NY: Cornell University Press.

Erzberger, Christian and Gerald Prein. 1997. "Optimal-Matching-Technik: Ein Analyseverfahren zur Vergleichbarkeit und Ordnung individuell differenter Lebensverläufe." [Optimal matching technique: an analytical process to compare and classify individual life courses] *ZUMA-Nachrichten* 40:52-80.

Everitt, Brian S. 1979. "Unresolved Problems in Cluster Analysis." *Biometrics* 35:169-81.

Forrest, John and Andrew Abbott. 1989. "The Optimal Matching Method for Studying Anthropological Sequence Data: An Introduction and Reliability Analysis." *Journal of Quantitative Anthropology* 1:151-70.

Giddens, Anthony, Mitchell Duneier, and Richard P. Applebaum. 2003. *Introduction to Sociology*. New York: W. W. Norton.

Giele, Janet Z. and Glen H. Elder. 1998. *Methods of Life Course Research: Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage.

Giuffre, Katherine A. 1999. "Sandpiles of Opportunity: Success in the Art World." *Social Forces* 77:815-32.

Goodman, Leo A. and William H. Kruskal. 1979. *Measures of Association for Cross Classification*. New York: Springer.

Grauer, Dan and Wen-Hsiung Li. 2000. *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer.

Halpin, Brendan and Tak Wing Chan. 1998. "Class Careers as Sequences: An Optimal Matching Analysis of Work-Life Histories." *American Sociological Review* 14:111-30.

Hartigan, John A. 1985. "Statistical Theory in Clustering." *Journal of Classification* 2:63-76.

Henikoff, Steven and Jorja G. Henikoff. 1992. "Amino Acid Substitution Matrices From Protein Blocks." *Proceedings of the National Academy of Sciences* 89:10915-19.

Höpflinger, François, Maria Charles, and Annelies Debrunner. 1991. *Familienleben und Berufsarbeit* (Family Life and Professional Work). Zurich, Switzerland: Seismo.

Hughey, Richard and Anders Krogh. 1996. "Hidden Markov Models for Sequence Analysis: Extension and Analysis of the Basic Method." *Computer Applications in Biological Science* 12:95-107.

Kohli, Martin. 1986. "The World We Forgot: A Historical Review of the Life Course." Pp. 271-303 in *Later Life: The Social Psychology of Aging*, edited by V. W. Marshall. London: Sage.

Krüger, Helga and René Levy. 2001. "Linking Life Courses, Work and the Family: Theorizing a Not So Visible Nexus Between Women and Men." *Canadian Journal of Sociology* 26:145-66.

Kruskal, Joseph. 1983. "An Overview of Sequence Comparison." Pp. 1-44 in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. Kruskal. Toronto, Canada: Addison-Wesley.

Lawrence, Charles E., Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, and John C. Wootton. 1993. "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment." *Science* 262:208-14.

Levine, Joel H. 2000. "But What Have You Done for Us Lately?" *Sociological Methods & Research* 29:34-40.

Levitt, Barbara and Clifford Nass. 1989. "The Lid on the Garbage Can: Institutional Constraints on Decision Making in the Technical Core of College-Text Publishers." *Administrative Science Quarterly* 34:190-207.

Levy, René. 1977. *Der Lebenslauf als Statusbiographie. Die weibliche Normalbiographie in makrosziologisher Perspektive*. [The life course as a sequence of statuses. The female standard biography in a macrosociological perpsective]. Stuttgart, Germany: Enke.

Levy, René, Jacques-Antoine Gauthier, and Eric Widmer. 2006. "Entre contraintes institutionnelle et domestique: Les parcours de vie masculins et féminins en Suisse." [Between institutional and domestic constraints: the life courses of women and men in Switzerland] *Revue canadienne de sociologie* 31:461-89.

Levy, René, Eric Widmer, and Jean Kellerhals. 2002. "Modern Family or Modernized Family Traditionalism? Master Status and the Gender Order in Switzerland." *Electronic Journal of Sociology* 6(4).

Manning, Gerard, David B. Whyte, Ricardo Martinez, Tony Hunter, and Sucha Sudarsanam. 2002. "The Protein Kinase Complement of the Human Genome." *Science* 298:1912-34.

Milligan, Glenn W. and Martha C. Cooper. 1985. "An Examination of Procedures for Determining the Number Clusters in a Dataset." *Psychometrika* 50:159-79.

Milligan, Glenn W. and Martha C. Cooper. 1987. "Methodology Review: Clustering Methods." *Applied Psychological Measurement* 11:329-54.

Moen, Phillis. 1985. "Continuities and Discontinuities in Women's Labor Force Activity." Pp. 113-55 in *Life Course Dynamics: Trajectories and Transitions, 1968-1980*, edited by G. H. Elder. Ithaca, NY: Cornell University Press.

Moen, Phillis. 2003. *It's About Time: Couples and Careers*. Ithaca, NY: Cornell University Press.

Moen, Phyllis and Yan Yu. 2000. "Effective Work/Life Strategies: Working Couples, Work Conditions, Gender, and Life Quality." *Social Problems* 47:291-326.

Mott, Frank L. 1978. *Women, Work and Family*. Lexington, MA: Lexington Books.

Müller, Tobias and Martin Vingron. 2000. "Modeling Amino Acid Replacement." *Journal of Computational Biology* 7:761-76.

Myrdal, Alva and Viola Klein. 1956. *Women's Two Roles: Home and Work*. London: Routledge.

Nargundkar, Satish and Timothy J. Olzer. 1998. "An Application of Cluster Analysis in the Financial Services Industry." Presented at the sixth annual conference of the South East SAS Users Group, September 13-15, Norfolk, VA.

Needleman, Saul B. and Christian D. Wunsch. 1970. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." *Journal of Molecular Biology* 48:443-53.

Ng, Pauline C., Jorja G. Henikoff, and Steven Henikoff. 2000. "PHAT: A Transmembrane-Specific Substitution Matrix. Predicted Hydrophobic and Transmembrane." *Bioinformatics* 16:760-66.

Notredame, Cédric, Philipp Bucher, Jacques-Antoine Gauthier, and Eric Widmer. 2005. *T-Coffee/saltt: User Guide and Reference Manual*. Lausanne: Swiss Institute of Bioinformatics. Retrieved from http://www.tcoffee.org/saltt.

Notredame, Cédric, Desmond G. Higgins, and Jaap Heringa. 2000. "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment." *Journal of Molecular Biology* 302:205-17.

Notredame, Cédric, Liisa Holm, and Desmond G. Higgins. 1998. "Coffee: An Objective Function for Multiple Sequence Alignments." *Bioinformatics* 14:407-22.

Olszak, Michael and Gilbert Ritschard. 1995. "The Behavior of Nominal and Ordinal Partial Association Measures." *Statistician* 44:195-212.

Pentland, Brian T., Malu Roldan, Ahmed A. Shabana, Louise L. Soe, and Sidne G. Ward. 1998. "Lexical and Sequential Variety in Organizational Processes." School of Labor and Industrial Relations, Michigan State University, East Lansing. Unpublished manuscript.

Punj, Girish and David W. Stewart. 1983. "Cluster Analysis in Marketing Research: Review and Suggestions for Application." *Journal of Marketing Research* 20:134-48.

Rapkin, Bruce D. and Douglas A. Luke. 1993. "Cluster Analysis in Community Research: Epistemology and Practice." *American Journal of Community Psychology* 21:247-77.

Rohwer, Götz and Ulrich Pötter. 2002. *TDA User's Manual*. Bochum, Germany: Ruhr Universität Bochum. Retrieved from http://www.stat.ruhr-uni-bochum.de/pub/tda/doc/tman63/tman-pdf.zip.

Rohwer, Götz and Heike Trappe. 1997. "Describing Life Courses. An Illustration Based on NLSY Data." Pp. 30 in *POLIS Project Conference*. Florence, Italy: European University Institute.

SAS Institute, Inc. 2004. *SAS/STAT User's Guide*. Cary, NC: Author.

Schaeper, Hildegard. 1999. "Erwerbesverläufe von Ausbildungsabsolventinnen und -Absolventen: Eine Anwendung der Optimal-Matching-Technik." [Employment history of girls and boys after completion of vocational education and training: an application of optimal matching technique]. Sonderforschungsbereich 186, Universität Bremen, Germany.

Scherer, Stefani. 2001. "Early Career Patterns: A Comparison of Great Britain and West Germany." *European Sociological Review* 17:119-44.

Sheridan, Jennifer T. 1997. "The Effects of the Determinants of Women's Movement Into and Out of Male Dominated Occupations on Occupational Sex Segregation." CDE Working Paper 97-07, Department of Sociology, University of Wisconsin, Madison.

Smith, Temple F. and Michael S. Waterman. 1981. "Identification of Common Molecular Subsequences." *Journal of Molecular Biology* 147:195-97.

Stovel, Katherine and Marc Bolan. 2004. "Residential Trajectories: Using Optimal Alignment to Reveal the Structure of Residential Mobility." *Sociological Methods & Research* 32:559-98.

Stovel, Katherine, Michael Savage, and Peter Bearman. 1996. "Ascription Into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970." *American Journal of Sociology* 102:358-99.

Thompson, Julie, Desmond G. Higgins, and Toby Gibson. 1994. "Clustal W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." *Nucleic Acids Research* 22:4673-80.

Turner, Jonathan H. 2001. "Sociological Theory Today." Pp. 1-17 in *Handbook of Sociological Theory*, edited by J. H. Turner. New York: Kluwer Academic.

Widmer, Eric, Jean Kellerhals, and René Levy. 2003. *Couples contemporains: Cohésion, régulation et conflits*. [Contemporary couples: cohesion, regulation, conflicts] Zürich: Seismo.

Widmer, Eric, Jean Kellerhals, and René Levy. 2004. "Quelle pluralisation des relations familiales?" [What pluralization of family relations]. *Revue française de sociologie* 45:37-67.

Widmer, Eric, René Levy, Alexandre Pollien, Raphael Hammer, and Jacques-Antoine Gauthier. 2003. "Entre standardisation, individualisation et sexuation: une analyse des trajectoires personnelles en Suisse" [Between standardization, individualization and gendering: an analysis of personal life courses in Switzerland] *Revue suisse de sociologie* 29:35-67

Wilson, W. Clarke. 1998. "Activity Pattern Analysis by Means of Sequence-Alignment Methods." *Environment and Planning A* 30:1017-38.

Wu, Lawrence L. 2000. "Some Comments on 'Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect.' " *Sociological Methods & Research* 29:41-64.

Yu, Yi-Kuo and Stefen F. Altschul. 2005. "The Construction of Amino Acid Substitution Matrices for the Comparison of Proteins With Non-Standard Compositions." *Bioinformatics* 21:902-11.

Jacques-Antoine Gauthier is a senior lecturer in sociology at the University of Lausanne and member of the Center for Life Course and Lifestyle Studies (Pavie). He has worked in the fields of health, addiction, and family sociology. His latest publications have appeared in the *Canadian Journal of Sociology, European Journal of Operational Research, and the Swiss Journal of Sociology*.

Eric D. Widmer is a professor of sociology at the University of Geneva, with an appointment at the Center for Life Course and Lifestyle Studies (Pavie). His long-term interests include life course research, family research, and social networks. His latest publications have appeared in the *Journal of Personal and Social Relationships*, *European Sociological Review*, and *Journal of Marriage and Family*.

Philipp Bucher is a group leader at the Swiss Institute for Experimental Cancer Research and a founding member of the Swiss Institute of Bioinformatics. His long-term interests include the development of algorithms for the analysis of molecular sequences and the application of these algorithms in various areas of biomedical research. His latest publications have appeared in *PLoS Computational Biology* and *Nucleic Acids Research*.

Cédric Notredame is a group leader at the Centre for Genomic Regulation in Barcelona (Spain) and a research investigator for the Centre National de la Recherche Scientifique (France). The focus of his work is the development and improvement of multiple sequence alignment algorithms. His latest publications have appeared in the *Journal of Molecular Biology* and *Nucleic Acid Research*. He is also the coauthor, with J. M. Claverie, of a popular introductory textbook in bioinformatics, *Bioinformatics for Dummies* (New York: Wiley, 2003).