A Study of Dynamic Time Warping for the Inference of Gene Regulatory Relationships

Matthias Böck, Constanze Schmitt, and Stefan Kramer

Technische Universität München,
Institut für Informatik Lehrstuhl I12 - Bioinformatik,
Boltzmannstr. 3, 85748 Garching b. München, Germany
{matthias.boeck,constanze.schmitt,stefan.kramer}@in.tum.de
http://wwwkramer.in.tum.de

Abstract. In this paper, we assess different variants of *Dynamic Time Warping* (DTW) for the inference of gene regulatory relationships. Apart from DTW on continuous time series, we present a novel angle-based discretization approach and a distance learning method that is combined with DTW to find new gene interactions. A positive influence of the distance optimization on the performance of the alignments of gene expression profiles could not yet be established. However, our results show that discretization can be important to the outcome of the alignments. The discretization is not only able to keep the important features of the time series, it is also able to perform better than regular DTW on the original data.

Keywords: Time series alignment, gene expression, Dynamic Time Warping, discretization

1 Introduction

The analysis of time series data is still one of the most challenging fields and occurs in many scientific disciplines. Steady state data can only give a snapshot of the actual dynamics while time series allow to study the processes over time and to capture the dependencies between the forces and protagonists. In this study we are focusing on gene expression data and how to infer the interactions and dependencies from it. We propose a slope based discretization of given microarray data and a new alignment approach, combining the ideas of *Dynamic Time Warping* (DTW) with *Stochastic Local Search* (SLS). Building of alignments of discretized profiles is supposed to be robust against noisy data and to overcome the assumption of strictly linear relationships between two interacting genes. A basic assumption for the alignment of time series is that co-regulated genes also show similar expression behavior over time and hence similar amplitudes which can be aligned with suitable transformations. Testing and evaluation of the approach has been done with one synthetic data set as well as four biological data sets.

2 Method

Dynamic time warping (DTW) was introduced in the 1960s [2] and has been intensively used for speech recognition and other fields, like handwriting recognition systems, gesture recognition, signal processing and gene expression time series clustering [1]. The basic idea of this unsupervised learning approach is that a suitable distance measure, which is most generally the Euclidean distance, allows the algorithm to stretch (or compress) the time and expression rate axis to find the most suitable fit of two given time series. The DTW algorithm will be described briefly in the following. Consider two given sequences $S = s_1, ..., s_n$ and $T = t_1, ..., t_m$ and a given distance function $\delta(s_i, t_j)$ with $1 \leq i \leq n$ and $1 \leq j \leq m$, DTW tries then to minimize with the given δ over all possible warping paths between the two given sequences based on the cumulative distance for each path. This is solved by a recursive dynamic programming approach for each $i \in [1, ..., n]$ and $j \in [1, ..., m]$:

$$DTW(i,j) = \begin{cases} 0 & \text{for } i = j = 0 \\ min \begin{cases} DTW_{i-1,j-1} + \delta(s_i, t_j) \\ DTW_{i-1,j} + \delta(s_i, t_j) \\ DTW_{i,j-1} + \delta(s_i, t_j) \end{cases} & \text{for } i, j > 0 \\ \infty & \text{otherwise} \end{cases} \tag{1}$$

$DTW[n, m]$ is the total distance $DTW(S, T)$ and can be calculated in $O(nm)$. The traceback through the matrix D gives the optimal warping of the aligned sequences. We use the symmetrical version of DTW which is supposed to perform better on equally sampled time series [12].

We present in the following the framework for a discretized sequence alignment approach focusing on the analysis of gene expression time series. In contrast to other existing methods, the approach deals also with anti-correlated time series and uses a supervised method to infer a data specific distance matrix for the alignment. The result is a scoring matrix for the pairwise distances between the measured genes.

We use four different gene expression time series of different lengths, for the evaluation. An overview of the networks and data sets will be given in the next section. All time series are centered around the x-axis by applying a z-score transformation to account for scale inconsistencies of the microarray experiments. We use cubic smoothing splines to interpolate the time series for missing values and smoothen out smaller fluctuations from experimental or biological noise.

The discretization of each time series for each gene is done according to the steepness of the expression change δ *exp* between two consecutive time steps. This is done by calculating the angle: α= **atan** δ *exp* $\cdot \frac{180}{\pi}$. The angles are then discretized into positive and negative (increasing or decreasing) integer values according to a predefined threshold. Defining the threshold is done by dividing the largest found angle for increases or decreases for each time series for a gene, into equally sized subsectors. Consider a maximum found angle of 180 degrees, which should be split into n subsectors, resulting a range of $\frac{180}{n}$ degrees each. Each of this sectors represents a possible range of angles for the increase or decrease between two consecutive time points and has assigned a discrete value. For n sectors the range of these values would be $[-\frac{n}{2}, -\frac{n}{2}+1, ..., \frac{n}{2}-1, \frac{n}{2}]$. To account for noise in the data, the two sectors which are neighboring the x-axis (one in the positive and one in the negative direction) are combined into one sector with the discrete value zero.

A crucial point for the quality of the alignments is the choice of a suitable distance matrix which defines the distances between the discretized values of the time series. This motivates our supervised approach to use a set of already known interacting genes I to infer the distance matrix δ. These gene pairs are chosen randomly from a given gold standard network along with a further randomly chosen set of not interacting genes N. The size of the latter is set if possible to twice the size of I. From this larger set N we resample between successive iterations of the distance calibration process new subsets to prevent accidentally chosen existing interaction partners between I and N genes to distort the result.

The resulting δ should minimize the distance for I and maximize the distance for N. Since DTW is not differentiable, we apply a combination of *Stochastic Local Search* (SLS) and simulated annealing for the stepwise improvement of δ. For a more detailed introduction to SLS, we refer to the work of Hoos and Stützle [4].

We imposed three constraints on the step-wise altering of the distance matrix δ to reduce the search space and to keep the basic distance structure between different bins of angles: $\delta(i,j) = 0$ for $i = j$, $\delta(i,j) = \delta(j,i)$ and $\delta(i,j) < \delta(i,j-1)$.

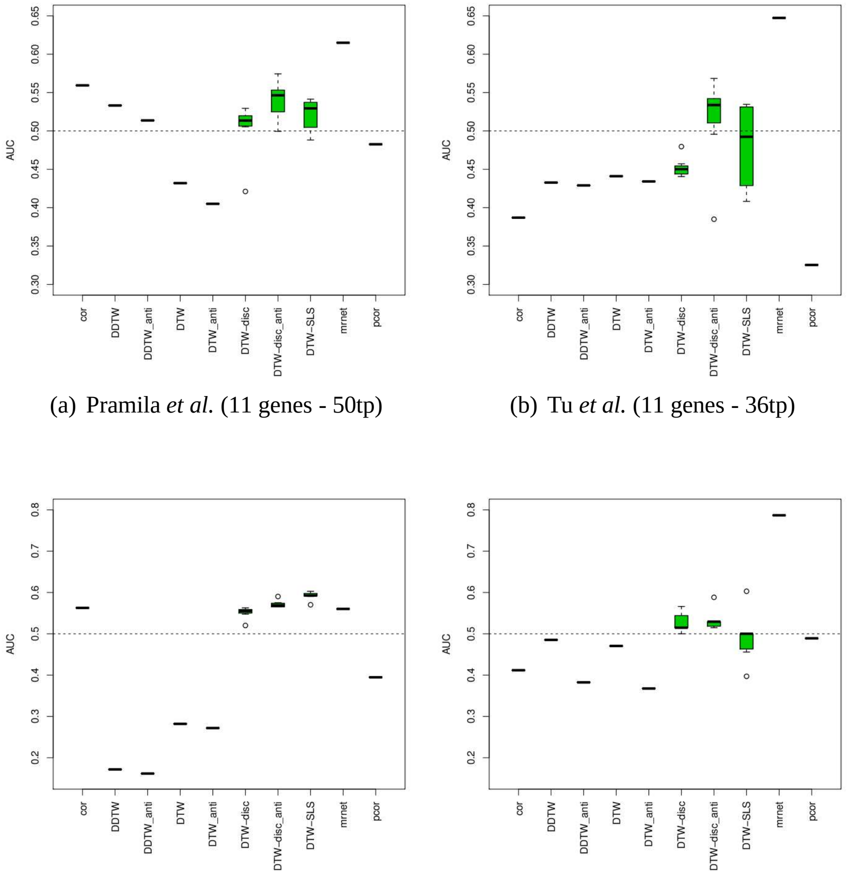
The resulting distance matrix is then used for the calculations of the alignments and the score defines the distance between each pair of genes. Additional alignments are done for each comparison with flipped signs for one of the time series to find anti-correlated pairs. All calculations were done in R except for the alignment matrix calculations, which were done for runtime efficiency in C.

3 Evaluation

The evaluation is done on five differently sized networks, a synthetic five gene network of yeast, called IRMA [3], the SOS signaling pathway in *E. coli* consisting of eight genes [10], a 11 cell cycle regulating network derived by Li *et al.* from the literature [7], and a full set of cell regulating genes, consisting of 1129 genes published by Rowicka *et al.* [11]. Gold standard and time series for the IRMA and SOS signaling pathway were taken from the R package *TDARACNE* [14] and consist of 16 and 14 measurements. The 11 cell cycle network by Li *et al.* as well as the suggested set of Rowicka *et al.* are tested with two time series experiments by Pramila *et al.* [9] and Tu *et al.* [13]. These sets follow several full cell cycles and include 50 and 36 time points. We left out genes of the large scale network which were not found in the experimental sets. This resulted in gene sets of size 961 and 944 for Pramila *et al.* and Tu *et al.*. As a benchmark network for the large scale cell cycle analysis, the protein-protein interaction network from the STRING database (v8.3) [5] is used. STRING calculates for each interaction a score based on the evidence from various sources like experiments, interaction databases or abstract text mining. We applied a cutoff of 0.8 to select only interactions with high confidence. It is clear that the PPI network is only able to cover part of the gene regulatory processes but still, observations on this level can provide insight into the performance of the methods. STRING is also considering pairs derived from co-expression analysis and might therefore be more suitable than other PPI databases. Self-regulations were excluded from all data sets

We compare the performance on the data sets to the results with simple correlation, partial correlation, MRNET (mutual information) [8], DTW and $DDTW$ (a modification of DTW which uses for the discretization the first derivative for each point) [6]. DTW_{disc} applies our discretization method with different numbers of sectors (n) and calculates the alignments with DTW. DTW_{SLS} additionally applies the distance calibration before the calculations. Methods ending with $\_anti$ also consider anti-correlated time series in the calculations. The evaluation is done based on ROC curves and the AUC. Interactions are undirected and hence only a two class problem considered, interaction predicted or not.

The results from the small networks in Figure 1 show that MRNET performs well even on shorter time series. Our methods perform only on the *E.coli* data better but are the second best performer on this task compared to the established methods. The discretized version performs in all cases, except for the Tu *et al.* data, better than guessing and outperforms DTW and $DDTW$, except for the Pramila *et al.* data, where $DDTW$ performs equally well. Including anti-



(a) Pramila *et al.* (11 genes - 50tp)

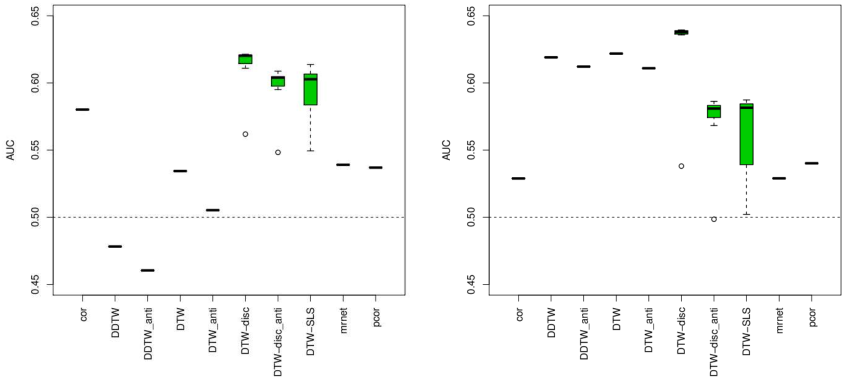
(b) Tu *et al.* (11 genes - 36tp)

(c) *E.coli* SOS (8 genes - 14tp)

(d) Synthetic yeast (5 genes - 16tp)

Fig. 1. Comparison of the performances on three small networks. The dotted line indicates guessing. The number of sectors ranges from 1 to 8. Knowledge size for calibration was set to 2. MR-NET performs best on all yeast data sets and only slightly worse than our proposed method on *E.coli*. The discretized version of DTW performs, except for case b), always better than guessing and best for the *E.coli* data set. Including anti-correlation improves in all cases the performance of DTW_{disc}. The other DTW versions perform quite differently on the data sets and in most cases even worse than guessing, especially in c).

correlation into the calculations improves in all cases of the discretized method the performance but has no positive effect for the regular DTW and $DDTW$. On the large scale network evaluation in Figure 2, the use of only correlated genes performs significantly better than with anti-correlation.



(a) Pramila *et al.* (961 genes - 50tp)

(b) Tu *et al.* (944 genes - 36tp)

Fig. 2. Comparison of the performances on a large scale network. The dotted line indicates guessing. Number of sectors ranges from 1 to 8. Knowledge size for calibration was set to 10. Gold Standard: STRING DB. Our approach performs significantly better in a) and only slightly worse in b). $DDTW$ and DTW perform best on the Tu *et al.* set but the influence of the anti-correlation is only small. DTW_{disc} performs much better in b) without the anti-correlation.

In general, the different DTW approaches perform better on the large scale data sets than correlation or MRNET, except in the case of regular DTW and $DDTW$ on the Pramila *et al.* data. The results of DTW_{disc} show that the discretization keeps the important features and performs well even with a small number of sectors. The approach of DTW_{SLS} seems, to this date, not to be able to improve the distance measure and achieves slightly smaller AUC values. The discretization method outperforms DTW and $DDTW$ on the Pramila *et al.* data and performs only slightly worse on the other data set.

4 Conclusion

In the paper, we investigated several variants of *Dynamic Time Warping* for the detection of gene regulatory relationships. While the supervised optimization of the distance matrix did not lead to improvements, a novel discretization approach seems, even with a small number of defined sectors, able to keep the main features and appears as a suitable qualitative transformation for time series alignments. On the biological data sets, our approach seems to be more stable compared to DTW and $DDTW$. In contrast to correlation-based methods, DTW is also able to infer the orientation of the time shift through the traceback and hence able to hint at possible causalities. We intend to make use of this information and further evaluate the robustness of the discretization method compared to DTW and $DDTW$.

References

1. J. Aach and G.M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 2001.
2. R. Bellman and R. Kalaba. On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2):1–9, November 1959.
3. I. Cantone, L. Marucci, F. Iorio, M. Ricci, V. Belcastro, M. Bansal, S. Santini, M. di Bernardo, D. di Bernardo, and M. Cosma. A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell*, (137):172–181, 2009.
4. H. H. Hoos and T Stützle. *Stochastic Local Search - Foundation and Applications*. Elsevier, 2005.
5. L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8 : a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl 1):D412–D416, 2009.
6. E. J. Keogh and M. J. Pazzani. Derivative Dynamic Time Warping. *Proceedings of the SIAM International Conference on Data Mining*, pages 1–11, 2001.
7. F. Li, T. Long, and C. Tang. The yeast cell-cycle network is robustly designed. *PNAS*, 101(14):4781–4786, 2004.
8. P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007.
9. T. Pramila, W. Wu, S. Miles, W.S. Noble, and L.L. Breeden. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes and Development*, 20(16):2266–2278, 2006.
10. M. Ronen, R. Rosenberg, B.I. Shraiman, and U. Alon. Assigning Numbers to the Arrows: Parameterizing a Gene Regulation Network by Using Accurate Expression Kinetics. *Proceedings of the National Academy of Sciences*, 99(16):10555–10560, 2002.
11. M. Rowicka, A. Kudlicki, B. P. Tu, and Z. Otwinowski. High-resolution timing of cell cycle-regulated gene expression. *Proceedings of the National Academy of Sciences*, 104(43):16892–16897, 2007.
12. H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49, 1978.
13. B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, 310:1152–1158, 2005.
14. P. Zoppoli, S. Morganella, and M. Ceccarelli. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics*, 11(1):154, 2010.