W482–W485 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl162

BTW: a web server for Boltzmann time warping of gene expression time series

F. Ferrè and P. Clote*

Department of Biology, Boston College, Chestnut Hill, MA USA

Received March 8, 2006; Accepted March 20, 2006

ABSTRACT

Dynamic time warping (DTW) is a well-known quadratic time algorithm to determine the smallest distance and optimal alignment between two numerical sequences, possibly of different length. Originally developed for speech recognition, this method has been used in data mining, medicine and bioinformatics. For gene expression time series data, time warping distance is arguably a more flexible tool to determine genes having similar temporal expression, hence possibly related biological function, than either Euclidean distance or correlation coefficient—especially since time warping accommodates sequences of different length. The BTW web server allows a user to upload two tab-separated text files A,B of gene expression data, each possibly having a different number of time intervals of different durations. BTW then computes time warping distance between each gene of A with each gene of B, using a recently developed symmetric algorithm which additionally computes the Boltzmann partition function and outputs Boltzmann pair probabilities. The Boltzmann pair probabilities, not available with any other existent software, suggest possible biological significance of certain positions in an optimal time warping alignment. Availability: http://bioinformatics.bc.edu/clotelab/BTW/.

INTRODUCTION

Dynamic time warping (DTW), described in the text of Kruskal and Liberman (1), is an algorithm to compute the optimal alignment of numerical sequences. Using dynamic programming (2), DTW was first introduced by Vintsyuk (3) and applied in speech recognition by Sakoe and Chiba (4). While most applications of DTW are in the area of speech recognition (5), DTW has recently been applied to the fields of data mining (6–8), medicine (electrocardiograms) (9) and bioinformatics (10,11). In bioinformatics, Aach and Church (10) implement some variants of classical DTW along with interpolative DTW, investigated robustness and statistical significance of time warping alignments, and analyzed published cell-cycle gene expression data of *Saccharomyces cerevisiae*. Very recently, Criel and Tsiporkova (11) describe their publicly available Java program, GenT χ Warper, which implements classical time warping and provides a user-friendly graphical user interface suitable for biological investigation of gene expression data.

DTW is a variant of sequence alignment for numerical sequences, as opposed to sequences of amino acids or nucleotides, where the analogue of a linear gap penalty in sequence alignment is played by time expansion or contraction. Despite the similarity with sequence alignment, DTW is subtly different and thus warrants a short presentation of details.

Let $a = a_1, \ldots, a_n$ be numerical values measured at times $0, \tau, 2\tau, \ldots, (n-1)\tau$, and let $b = b_1, \ldots, b_m$ be numerical values measured at times $0, \mu, 2\mu, \ldots, (m-1)\mu$. For each $1 \leq i \leq n$, $1 \leq j \leq m$, define

$$D_{i,j} = \min\begin{cases} D_{i-1,j-1} + \frac{\tau+\mu}{2} \cdot |a_i - b_j| \\ D_{i-1,j} + \frac{\tau}{2} \cdot |a_i - b_j| \\ D_{i,j-1} + \frac{\mu}{2} \cdot |a_i - b_j| \end{cases}$$

With this notation, (classical) time warping distance between a and b is defined to be $D_{n,m}$; the optimal alignment is obtained using tracebacks. Clearly the computation of time warping distance can be performed in quadratic time using quadratic memory resources. Using the method of Hirschberg (12), it is possible to reduce memory resources to a linear factor; however, given the current number of time points in gene expression data, this additional complication is unwarranted.

As in sequence alignment, time warping can alternatively be viewed as a method to determine the minimum cost path, which proceeds in a left-to-right and bottom-to-top manner from the point (1,1) to the point (n,m), such that each path edge is one step in the horizontal, vertical or diagonal direction. See Figure 1 for an example of the optimal path graph between

*To whom correspondence should be addressed at Departments of Biology and Computer Science (courtesy appointment), Boston College, Chestnut Hill, MA USA. Tel: +1 617 552 1332; Fax: +1 617 552 2011; Email: clote@bc.edu

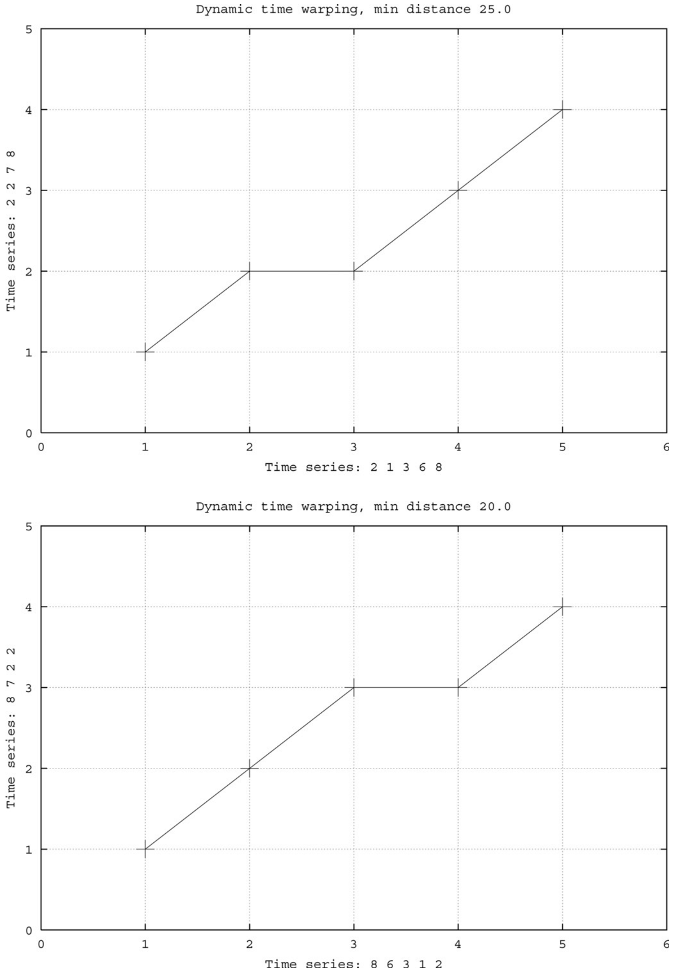


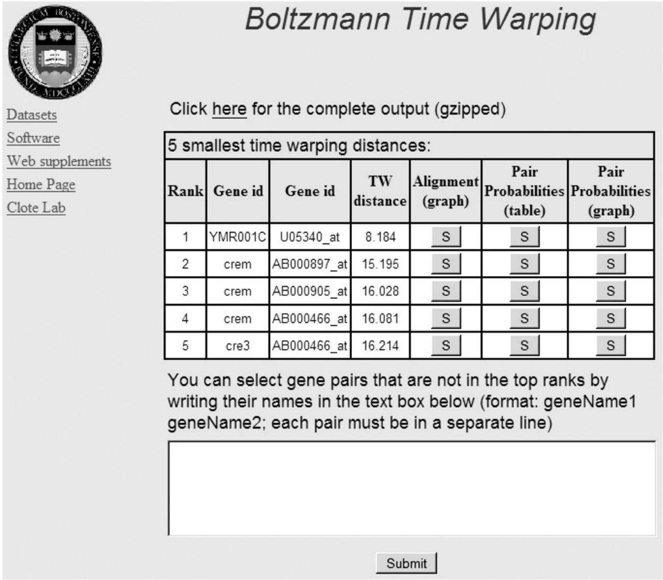
Figure 1. Path graph for classic DTW for (toy) sequences $a = (a_1,a_2,a_3,a_4,a_5) = (2,1,3,6,8)$ and $(b_1,b_2,b_3,b_4) = (2,2,7,8)$ of unequal length. Here $|a_i - b_j|$ is Euclidean distance, and time expansion/compression intervals are $\tau = 10 = \mu$. (Left) Optimal path graph with distance 25.0 for aligning a with b. (Right) Optimal path graph with distance 20.0 for aligning the reversal of a with the reversal of b. Note that classic DTW is not symmetric, unlike the situation for sequence alignment algorithm of Needleman and Wunsch (20).

two small numerical sequences, and notice that unlike the case for sequence alignment, DTW is not symmetric—i.e. time warping distance between sequences a,b is not necessarily equal to that of the reversal of a with the reversal of b. A Sakoe–Chiba band (4) is used to define a warping window that constrains the path close to the diagonal, restricting the portion of the matrix that the path is allowed to visit to

$$|i - j \cdot \frac{n}{m}| < p \cdot n$$

given a parameter p between 0 and 1.

In (13), Clote and Straubhaar describe a new variant of time warping, which is proved to be symmetric in the sense just described. The paper (13) includes a quadratic time computation of the Boltzmann forward partition function $FZ_{i,j} = \sum \exp(-D_{i,j}/RT)$, where the sum is over all time warpings of $a_1, \ldots, a_i$ with $b_1, \ldots, b_j$, $D_{i,j}$ is the (new, symmetric and modified) time warping distance between $a_1, \ldots, a_i$ and $b_1, \ldots, b_j$, R is the universal gas constant and T is absolute temperature. In the context of time warping, T has no physical significance and can be chosen arbitrarily. Symmetry allows the unambiguous computation of the backward partition function BZ_{ij}, where the sum is over all time warpings of $a_i, \ldots, a_n$ with $b_j, \ldots, b_m$. Together we obtain the Boltzmann pair probability $Pr[a_i,b_j]$ that a_i is aligned with b_j, defined by



Boltzmann Time Warping

Click here for the complete output (gzipped)

5 smallest time warping distances:

Rank	Gene id	Gene id	TW distance	Alignment (graph)	Pair Probabilities (table)	Pair Probabilities (graph)
1	YMR001C	U05340_at	8.184	S	S	S
2	crem	AB000897_at	15.195	S	S	S
3	crem	AB000905_at	16.028	S	S	S
4	crem	AB000466_at	16.081	S	S	S
5	cre3	AB000466_at	16.214	S	S	S

You can select gene pairs that are not in the top ranks by writing their names in the text box below (format: geneName1 geneName2; each pair must be in a separate line)

Figure 2. Output from BTW: top ranking time warping distance for genes from file A with those from file B. Small time warping distance could indicate a related biological function, as determined in the Gene Ontology. For each gene pair, the user can retrieve a graphical description of the optimal alignment, as well as text and graphical output for the Boltzmann pair probabilities $Pr[a_i,b_j]$. Optimal alignment and pair probabilities graphical output is shown in Figure 3.

$$Pr[a_i, b_j] = \frac{FZ_{i-1,j-1} \cdot \exp(-\frac{(\tau+\mu)\cdot|a_i - b_j|}{2RT}) \cdot BZ_{i+1,j+1}}{FZ_{1,n}}$$

This expression is a slight simplification—see (13) for more details. Boltzmann pair probabilities indicate possible biological significance for certain alignment positions in time warping of gene expression data. For instance, it could be that the alignment of expression values in cell-cycle phases G_1 and S is more significant than those in phases G_2 and M. Indeed, in the context of sequence alignment, (14) has shown that Boltzmann pair probabilities do indicate biological significance of certain positions in sequence alignments; see also (15). Although currently available gene expression time series data include only a small number of time points, future datasets are certain to include more time points, as costs decrease and accuracy is improved. Hence, our BTW web server should prove increasingly useful in genomics research.

WEB SERVER

Figure 2 displays a screen shot of the BTW web server which computes the optimal symmetric time warping distance between two numerical sequences, and allows the user to obtain the Boltzmann pair probabilities $Pr[a_i,b_j]$ that a_i and b_j are aligned. Additionally graphical output is available, as depicted in Figure 3.

The BTW web server currently allows a user to upload two tab-separated text files, A and B, each containing gene

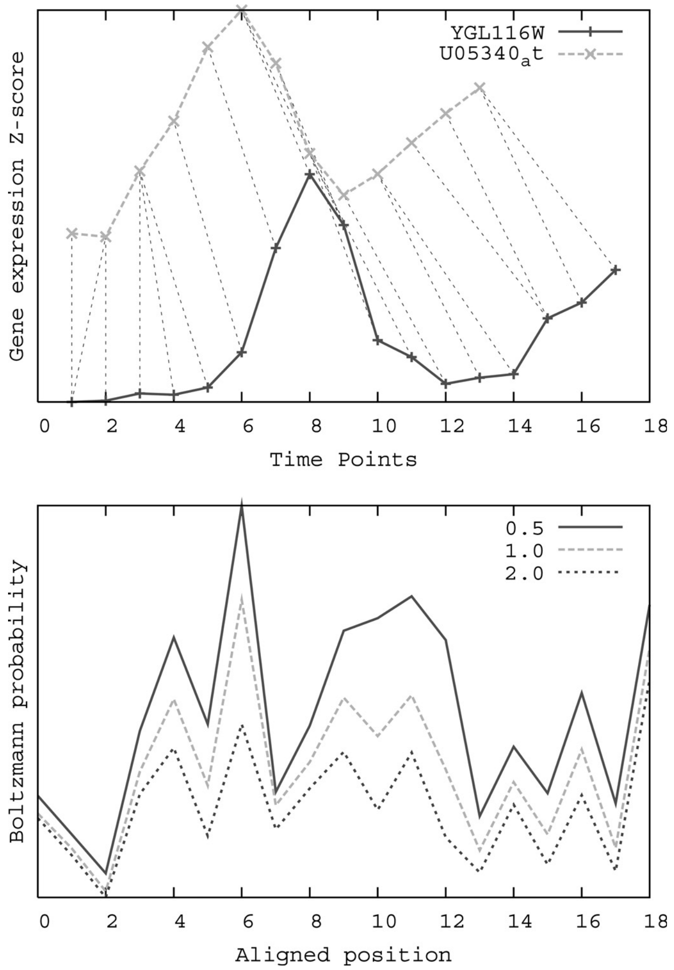


Figure 3. Optimal alignment (upper panel) and Boltzmann pair probabilities (lower panel) for alignment positions in optimal time warping between yeast gene YGL116W (17 time points) and human gene U05340 (13 time points). Boltzmann probabilities are computed with $k = 1$ and for the values 0.5, 1.0 and 2.0 of T. With increasing values of T, the Boltzmann probability values decrease; this enhances the depiction of presumed biological significance of aligned positions in this optimal time warping.

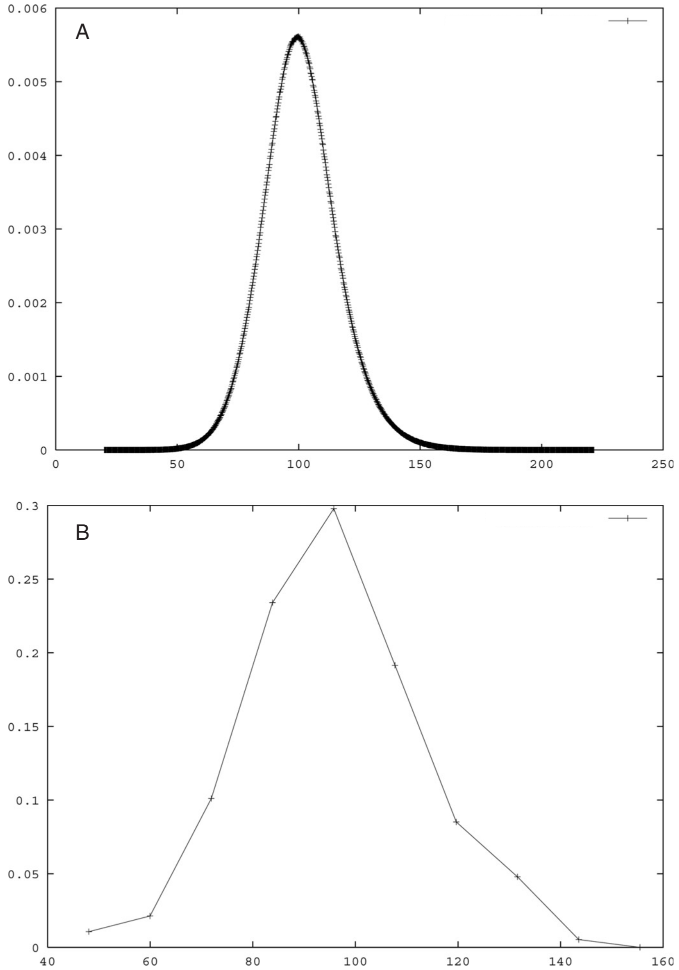


Figure 4. (**A**) Histogram of (symmetric) time warping distances between each yeast and human gene, using gene expression data from Cho *et al.* (16,17); mean $\mu = 100.91$, SD $\sigma = 15.647$, max $= 221.74$, min $= 19.99$. (**B**) Histogram of (symmetric) time warping distances between each yeast and human gene from a sample of 188 homologous pairs of yeast/human genes; mean $\mu = 101.73$, SD $\sigma = 17.75$, max $= 167.41$, min $= 48.03$. Homologous pairs determined using NCBI HomoloGene (J. Straubhaar personal communication).

expression time series data. The first line is required to contain the time points, measured in minutes—for instance, file A might contain values $t_1, \ldots, t_n$ equal to $0, 10, 20, \ldots, 160$ as in Cho's expression data for *S.cerevisiae* (16), and file B might contain values $t'_1, \ldots, t'_m$ equal to $0, 120, 240, \ldots, 1440$ as in Cho's expression data for *Homo sapiens* (17). Time intervals need not be constant; for instance, a time series might consist of values measured at times 0, 20, 40, 80, 160. The user may enter constants c resp. d, set by default to 1, in order to allow scaling between τ and μ; i.e. $\tau' = c \cdot \tau$ resp. $\mu' = \mathrm{d} \cdot \mu$. For instance, since approximately two cell-cycles for *S.cerevisiae* occur in 160 min for the data of (16) with intervals of 10 min, and since approximately two cell-cycles for *H.sapiens* occur in 1440 min (24 h) or the data of (17) with intervals of 120 min, there are 16 non-zero time points for yeast compared with 12 time points for human. Hence one could take $c = 1/10$, $d = 4/3 \cdot 1/120$ or alternatively $c = 1$, $d = 4/3 \cdot 1/12$. The leftmost column of both A and B contains distinct gene names.

Every subsequent line of file A resp. B is required to contain normalized log expression values, with no missing data—the user is assumed to complete missing data by using interpolation, splines (18) or another method. The BTW web server computes the (symmetric) time warping distance between each gene of A and each gene of B, where due to time and space constraints, file A resp. B has an upper bound of 100 resp. 10 000 genes. Time warping scores for gene pairs are sorted by increasing distance, or available in lexicographic order of gene pairs. For up to 100 user-specified gene pairs, Boltzmann pair probabilities are available in both text and graphical format, the latter depicted in Figure 3. Since Clote and Straubhaar (13) proved the symmetry of one of the four variants of DTW in (10), the BTW web server allows the user to choose either the algorithm of Clote or that of Aach. [After seeing a preprint of (13) Aach conjectured that his DTW program genewarp with flag -a 2 is symmetric. Aach's conjecture is proved in (13)].

Current hardware supporting the BTW web server consists of a Beowulf-style cluster comprising 6 Dell 1650, 2 × 1300 MHz Pentium III, 2 GB RAM with 4 Apple XServe, 2 × 1333 MHz G4, 2 GB RAM and finally 12 Dell 1850, 2 × 2800 MHz Xeon EM64T, 2 GB RAM. Interconnect is 1 Gbit Ethernet. Pentium III nodes are running 32-bit CentOS 4.2, Xeon EM64T nodes are running 64-bit CentOS 4.2 and G4 nodes are running MacOS 10.2.8.

DISCUSSION

A histogram containing 1000 classes, produced over 40 million [6601 yeast × 7077 human (16,17)] time warping distances between each yeast and each human gene is depicted in Figure 4A. This figure suggests that time warping distances could follow an extreme value distribution, known by (19) to be the distribution of BLAST hits. From the method of moments, we fit the histogram of Figure 4A to an extreme value distribution with cumulative distribution function $CDF(x) = e^{-e^{-(x-93.87)/12.20}}$. This would allow us to compute a P-value for significance of human genes, whose time warping distance with a given yeast gene is less than a fixed threshold.

ACKNOWLEDGEMENTS

Research was partially supported by National Science Foundation grant DBI-0543506. Funding to pay the Open Access publication charges for this article was provided by the National Science Foundation (NSF) with grant DBI-0543506.

Conflict of interest statement. None declared.

REFERENCES

1. Kruskal,J.B. and Liberman,M. (1999) The symmetric time-warping problem: from continuous to discrete. In Kruskal,J.B. and Sankoff,D. (eds), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. CSLI Publications, Stanford, pp. 125–161.
2. Bellman,R.E. and Dreyfus,S.E. (1959) Functional approximations and dynamic programming. *Math. Tables and other Aids Comp.*, **13**, 247–251.
3. Vintsyuk,T.K. (1968) Speech discrimination by dynamic programming. *Kibernetka (Cybernetics)*, **4**, 81–88.
4. Sakoe,H. and Chiba,S. (1978) Dynamic programming algorihtm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, Signal Proc.*, **26**, 43–49.
5. Rabiner,L. and Juang,B. (1993) *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ.
6. Keogh,E. and Pazzani,M. (1998) An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In *Proceedings of the Fourth International Conference of Knowledge Discovery and Data Mining (KDD'98)*. AAAI Press, pp. 239–241.
7. Keogh,E. (2003) Efficiently finding arbitrarily scaled patterns in massive time series databases. In Lavrac,N., Gamberger,D., Todorovski,L. and Blockee,H. (eds), *Knowledge Discovery in Databases: PKDD 2003, 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Springer Lecture Notes in Computer Science 2838, Springer, Cavtat-Dubrovnik, Croatia, pp. 253–265.
8. Keogh,E. and Ratanamahatana,C. (2005) Exact indexing of dynamic time warping. *Knowledge and Information Systems*, **7**, 358–386.
9. Caiani,E.G., Porta,A., Baselli,G., Turiel,M., Muzzupappa,S., Pagani,M., Malliani,A. and Cerutti,S. (2002) Analysis of cardiac left-ventricular volume based on time warping averaging. *Med. Biol. Eng. Comput.*, **40**, 225–233.
10. Aach,J. and Church,G.M. (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics*, **17**, 495–508.
11. Criel,J. and Tsiporkova,E. (2006) Gene Time E{chi} pression Warper: a tool for alignment, template matching and visualization of gene expression time series. *Bioinformatics*, **22**, 251–252.
12. Hirschberg,D. (1975) A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, **18**, 341–343.
13. Clote,P. and Straubhaar,J. (2005) Symmetric time warping, Boltzmann pair probabilities and functional genomics. *J. Math. Biol.* In Press.
14. Vingron,M. and Argos,P. (1990) Determination of reliable regions in protein sequence alignments. *Protein Eng.*, **3**, 565–569.
15. Muckstein,U., Hofacker,I.L. and Stadler,P.F. (2002) Stochastic pairwise alignments. *Bioinformatics*, **18**, S153–S160.
16. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
17. Cho,R.J., Huang,M., Campbell,M.J., Dong,H., Steinmetz,L., Sapinoso,L., Hampton,G., Elledge,S.J., Davis,R.W. and Lockhart,D.J. (2001) Transcriptional regulation and function during the human cell cycle. *Nature Genet.*, **27**, 48–54.
18. Bar-Joseph,Z., Gerber,G.K., Gifford,D.K., Jaakkola,T.S. and Simon,I. (2003) Continuous representations of time-series gene expression data. *J. Comput. Biol.*, **10**, 341–356.
19. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
20. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.