Web supplement
"BTW: A web server for Boltzmann time warping of gene expression time series"

F. Ferrè and P. Clote[*]
Department of Biology
Higgins 355, Boston College, Chestnut Hill, MA 02467, USA
Tel: 617 552 1332, Fax: 617 552 2011, {ferref,clote}@bc.edu.

Abstract

Here we address the question of numerical precision in the Boltzmann partition function computation for time warping alignments of two time series of gene expression data. This point was raised by an anonymous referee, whom we would like to thank.

1 Precision

We assume familiarity with the notions of time warping, partition function, and results of [1, 4]. In particular, the *partition function* $Z^T(a,b)$ for time warping alignments between sequences $a = a_1, \ldots, a_n$ and $b = b_1, \ldots, b_m$ is defined to be

$$Z^T(a,b) = \sum e^{-\rho(\mathcal{A})/RT} \tag{1}$$

where the sum is over all possible time warpings $\mathcal{A}$ of sequences a and b, $\rho(\mathcal{A})$ is the (symmetric) time warping distance between a and b for the alignment $\mathcal{A}$, R is the universal gas constant (Boltzmann constant times Avogadro's number), and T is absolute temperature. Our time warping server, *BTW*

[*] Corresponding author.

Sequences	Mean	Stdev	Min	Max
Y2Y	111.65	20.02	0.56	270.23
H2Y	104.61	15.71	21.68	223.47
H2H	89.70	180.42	12.57	184.70

Table 1: Statistics on time warping distances for yeast and human data sets of R. Cho et al. [2, 3]. The rubric Y2Y denotes the measurements when time warping yeast with yeast (all against all). Similarly H2Y denotes human with yeast and H2H denots human with human.

computes the partition function Z to 3 decimal places. Given current data, this is more than justified by the following considerations.

In computing the time warping distance for Cho's yeast and human data sets [2, 3], we obtain the statistics presented in Table 1. Temperature T and universal gas constant R have no physical meaning in the context of time warping alignments. For this reason, we set the value of R to be 1. Formal temperature T is a user-defined integer, but is generally in the half-open interval $(0, 300]$. In the computations below, we assume that $T = 1$.

By Theorem 9.1 on page 187 of [5], the number of alignments of two sequences of lengths n and m, respectively, is given by the recursion relation

$$f(n,m) = f(n-1,m) + f(n,m-1) + f(n-1,m-1)$$

and asymptotically satisfies

$$f(n,n) \sim (1+\sqrt{2})^{2n+1} n^{-1/2}.$$

This asymptotic limit requires values of n which far exceed the lengths of current time series, so using simple dynamic programming, the exact number of alignments is computed in the Python program `numPrecisionTimeWarp.py`, available on the server.

Gene expression values are usually reported as logarithms of the measured expression values, or in the case of microarrays, as logarithm of the relative expression values. When using our software BTW, it is assumed that all data has been *normalized*, so that values have mean of 0 and standard deviation of

1. To avoid any ambiguity, from this point on, we use the term *log expression values* to denote the normalized values.[1]

Suppose that the reported log expression values are correct to within ϵ accuracy. In the worst case scenario, the error in computing time warping distance between two length n time series of log expression values is approximately bounded by $n\epsilon$. By definition,

$$Z^T(n,m) = \sum_{\mathcal{A}} e^{-\rho(\mathcal{A})/RT}$$

where the sum is taken over all possible time warping alignments $\mathcal{A}$, and where $\rho(\mathcal{A})$ denotes the time warping distance for the alignment $\mathcal{A}$. Recall that $n\epsilon$ is a worst-case upper bound for the error between the time warping distance for the correct sequences $a = a_1, \dots, a_m$ and $b = b_1, \dots, b_m$ versus that of the erroneously measured sequences $a' = a'_1, \dots, a'_m$ and $b' = b'_1, \dots, b'_m$.[2]

Let ΔZ denote the error in the partition function calculation $Z^T(n,m)$ for time warping sequences of lengths n and m, respectively. Then

$$\begin{array}{rcl}
\Delta Z^T(n,m) & = & Z^T(a,b) - Z^T(a',b') \\
& \le & \sum_{\mathcal{A}} e^{-(\rho+\Delta\rho)/RT} - e^{-(\rho/RT} \\
& = & \sum_{\mathcal{A}} |\frac{e^{-(\rho+\Delta\rho)/RT} - e^{-(\rho/RT}}{\Delta\rho}| \cdot \Delta\rho \\
& \approx & f(n,m) \cdot |\frac{d}{d\rho}\left(e^{-\rho/RT}\right)| \cdot \Delta\rho \\
& \le & f(n,m) \cdot \frac{\Delta\rho}{RTe^{\rho/RT}}
\end{array}$$

[1] Before normalization, values can be quite large. For instance, orf YOL121C of gene RPS16Bex1f in yeast, from R. Cho et al. [2, 3], has the 17 time points given by: 2143, 2474, 2889, 2216, 2542, 1705, 1703, 1970, 1775, 6235, 3682, 2281, 1872, 2721, 2583, 3488, 3205. After normalization, these values are given by: -0.500695247312, -0.189482151474, 0.200709494366, -0.432059126477, -0.125547134806, -0.912511972318, -0.914392413985, -0.663353451481, -0.846696513984, 3.34668840275, 0.946304615211, -0.37094477231, -0.755495093149, 0.0427523943634, -0.0869980806386, 0.763901773541, 0.497819277704. From this perspective, it is not unreasonable to assume that the error ϵ in the normalized data is at most 0.1 or even 0.001. Computations are reported below for these two values of ϵ.

[2] We report two computations, where the second more generously assumes that the error bound is ϵ rather than $n\epsilon$.

n	m	$f(n,m)$	ϵ	ρ	Err	ΔZ
13	17	30845044155	0.001	30	$n\epsilon$	5.77272587319e-05
13	17	30845044155	0.001	30	ϵ	2.8863629366e-06
13	17	30845044155	0.1	30	$n\epsilon$	0.00577
13	17	30845044155	0.1	30	ϵ	0.0002886
20	20	260543813797441	0.001	30	$n\epsilon$	0.4876
20	20	260543813797441	0.001	30	ϵ	0.0244
20	20	260543813797441	0.1	30	$n\epsilon$	48.76
20	20	260543813797441	0.1	30	ϵ	2.438

Table 2: Approximate error bound for partition function computation produced by program `numPrecisionTimeWarp.py`. Given time series $a = a_1, \ldots, a_n$ and $b = b_1, \ldots, b_m$ of lengths n resp. m, we assume that the measured time series $a' = a'_1, \ldots, a'_n$ and $b' = b'_1, \ldots, b'_m$ has error bound ϵ; i.e $|a_i - a'_i|, |b_j - b'_j| < \epsilon$, for all i, j. Let $D_{i,j}(a, b)$ denote the time warping distance between prefixes $a_1, \ldots, a_i$ and $b_1, \ldots, b_j$. Similarly, let $D_{i,j}(a', b')$ denote the time warping distance between prefixes $a'_1, \ldots, a'_i$ and $b'_1, \ldots, b'_j$. We assume that error 'Err' in time warping satisfies $|D_{n,m}(a, b) - D_{n,m}(a', b')| \leq Err$, and that either Err $= n \cdot \epsilon$, or Err $= \epsilon$. Value $f(n, m)$ is the number of alignments of a with b, and the approximate error bound in partition function computation, ΔZ, is defined in equation (2). We considered the values $17, 13$ because 17 resp 13 is the number of time points in yeast resp. human time series data of Cho et al. [2, 3].

It follows that

$$\Delta Z = \frac{f(n,m) \cdot \Delta\rho}{RTe^{\rho/RT}} \tag{2}$$

is an approximate upper bound for error in the partition function computation introduced because of an error of ϵ in the log expression values. Using the program `numPrecisionTimeWarp.py`, available on the server, we performed some omputations for the yeast and human data sets of R. Cho et al.

Table 2 gives some bounds of the approximate error bound ΔZ in the partition function computation, where this error is induced by measurement error ϵ per data point in the time series. When time warping human and yeast data, there are $30,845,044,155 \approx 3.08 \times 10^{10}$ many pairwise alignments, and from Table 1, the mean value of ρ is 104.61, with standard deviation 15.71, while the minimum value of ρ is 21.68. For this reason, in Table 2, we took $n = 30$. For the range of data that the user is likely to enter in our web server BTW, we feel that 3-place accuracy is justified.

References

[1] P. Clote and J. Straubhaar. Symmetric time warping, Boltzmann pair probabilities and functional genomics. *J. Math. Biol.*, 2006. in press.

[2] R. Cho et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[3] R. Cho et al. Transcriptional regulation and function during the human cell cycle. *Nature Genetics*, 27:48–54, 2001.

[4] F. Ferrè and P. Clote. BTW: A web server for Boltzmann time warping of gene expression time series. 2006. submitted.

[5] M.S. Waterman. *Introduction to Computational Molecular Biology: Maps, sequences and genomes.* Chapman & Hall – CRC Press, 1995.