BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 3 2014, pages 377–383
doi:10.1093/bioinformatics/btt692

Gene expression

Advance Access publication November 27, 2013

DDGni: Dynamic delay gene-network inference from high-temporal data using gapped local alignment

Hari Krishna Yalamanchili[1,2], Bin Yan[3], Mulin Jun Li[1,2], Jing Qin[1,2], Zhongying Zhao[3], Francis Y.L. Chin[4] and Junwen Wang[1,2,5,*]

[1]Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, [2]Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, [3]Department of Biology, Hong Kong Baptist University, Kowloon, [4]Department of Computer Science, Faculty of Engineering and [5]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Inferring gene-regulatory networks is very crucial in decoding various complex mechanisms in biological systems. Synthesis of a fully functional transcriptional factor/protein from DNA involves series of reactions, leading to a delay in gene regulation. The complexity increases with the dynamic delay induced by other small molecules involved in gene regulation, and noisy cellular environment. The dynamic delay in gene regulation is quite evident in high-temporal live cell lineage-imaging data. Although a number of gene-network-inference methods are proposed, most of them ignore the associated dynamic time delay.

Results: Here, we propose DDGni (dynamic delay gene-network inference), a novel gene-network-inference algorithm based on the gapped local alignment of gene-expression profiles. The local alignment can detect short-term gene regulations, that are usually overlooked by traditional correlation and mutual Information based methods. DDGni uses 'gaps' to handle the dynamic delay and non-uniform sampling frequency in high-temporal data, like live cell imaging data. Our algorithm is evaluated on synthetic and yeast cell cycle data, and *Caenorhabditis elegans* live cell imaging data against other prominent methods. The area under the curve of our method is significantly higher when compared to other methods on all three datasets.

Availability: The program, datasets and supplementary files are available at http://www.jjwanglab.org/DDGni/.

Contact: junwen@hku.hk

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 22, 2013; revised on November 14, 2013; accepted on November 19, 2013

1 INTRODUCTION

Biological systems involve collaborations at various levels, from the atomic interactions to more complex ecosystems. Network formulation will enable a better understanding of such complex systems. In particular, the interplay between genes and the networks they constitute are of great interest to many biologists. Gene networks can help us to comprehend the cause, prognosis and to prioritize drug targets of various diseases (Madhamshettiwar *et al.*, 2012). The spatial and temporal dynamics in gene-expression patterns can reveal regulatory pathways (networks) and thereby help us to understand various underlying mechanisms (Davidson and Levine, 2005). Recent advances in live cell imaging techniques have enabled continuous documentation of cell divisions and quantitative measurement of gene expressions for each cell throughout embryogenesis (Murray *et al.*, 2006). Inferring regulatory networks from such data will help to decode various regulatory mechanisms involved in tissue differentiation and embryonic development.

Reconstruction of gene-regulatory networks (GRN) accurately from high-temporal data is fundamental but still remains a challenge. Gene regulations are not spontaneous (Josic *et al.*, 2011); various processes are involved in producing fully functional and measurable concentrations of transcriptional factors/proteins. Each process takes time, leading to the delay in gene regulation (Zhu *et al.*, 2007). Furthermore, cellular mRNA and protein concentrations are substantially influenced by noisy cellular environment and other small molecules involved in gene regulation (Bratsun *et al.*, 2005). These fluctuations (noise) can lead to the dynamic increase or decrease in delay, during transcriptional regulation as illustrated in Figure 1a (Blake *et al.*, 2003). The dynamic delay in gene regulation is quite evident in high-temporal live cell lineage-imaging data (Murray *et al.*, 2006).

Quite a number of gene-network-inference methods are proposed. However, their performance varies on different datasets with respective limitations. For instance, correlation based methods perform better in predicting linear relationships whereas, information theoretic (Mutual Information, MI) based methods are better for non-linear relationships. Network topology is also critical, as some methods are more suitable for Erdos–Renyi random topology and others for standard small world scale free networks (Stolovitzky *et al.*, 2009). Most of the methods perform better on steady state data as compared to temporal data (Marbach *et al.*, 2010), due to the fluctuations in magnitude and the delay in gene expression.

Current GRN-inference methods mainly focus on transcriptional events. However, most of them ignore the associated dynamic time delay. Cross-correlation-based methods can identify

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

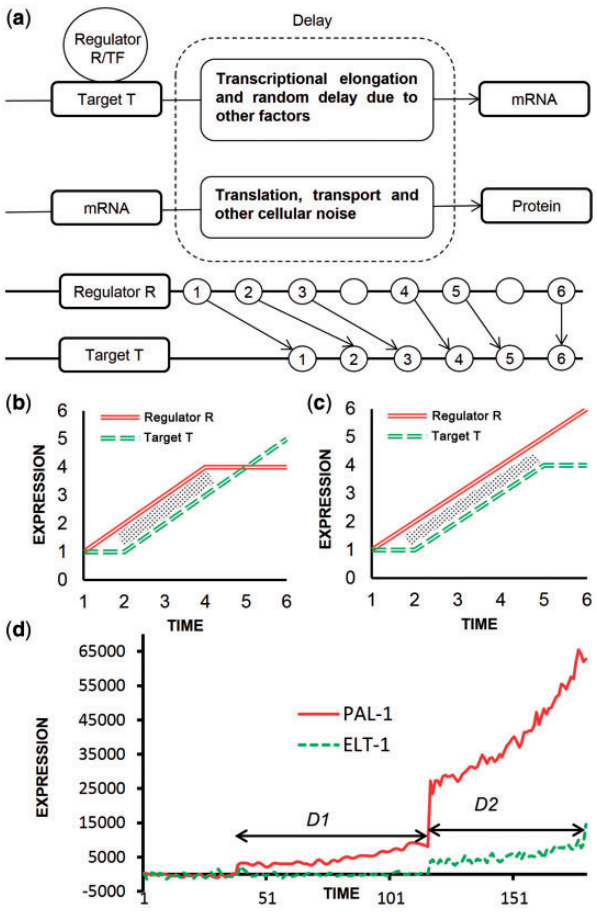

Fig. 1. Illustration of dynamic time delay in gene regulation (**a**) Origin of delay: cascade of reactions between the binding of regulator/transcription factor and the detection of matured protein. Noisy cellular environment and other small factors add to the transcriptional and translational delay. (**b** and **c**) Model gene-expression patterns with local correlations (shaded regions are windows of regulation). (**d**) Positive regulation between the genes PAL-1-1 and ELT-1 in the C lineage of *C.elegans* with varying delay ($D1 \neq D2$)

delayed correlations. Nonetheless, to find the maximal delayed correlation, one needs to compute correlations for all the *N-1* possible delays between a regulator and its target, where *N* is the number of time points in the gene-expression data (Rhudy *et al.*, 2010); this makes it computationally expensive for high-temporal data. Moreover, it cannot be applied for identification of non-linear relationships. On the other hand, although the time-delayed MI based methods can predict non-linear relationships, yet there is a need to compute MI for all the *N-1* possible delays. MI assumes long sampling intervals and statistical independence between time points (Huang *et al.*, 2010). However, this assumption is not valid for continuous live-cell-imaging data (Murray *et al.*, 2006) as the expression at time *t* is dependent on its previous time stamp and it is imperative to consider such dependencies.

Time Delay-ARACNE (Zoppoli *et al.*, 2010) is a three-step information-theoretic-based method. Starting with the gene-expression change-point analysis, it is followed by network construction and then pruning. However, being an information theoretic method, it inherits limitations such as statistical independence between time points and long sampling intervals (Huang *et al.*, 2010).

GeneReg (Huang *et al.*, 2010) is a regression-based method similar to Ordinary Differential Equations (ODEs). It uses a linear model with time delay and regulation coefficient as its parameters. This assumes a constant delay between the regulator and its target. In reality however, the delay is dynamic and is influenced by other small molecules involved in gene regulation and noisy cell environment (Bratsun *et al.*, 2005).

Recently, Dynamic Time Wrapping (DTW) based similarity measures are also used to infer time delay gene networks (Aach and Church, 2001; Lee *et al.*, 2012; Riccadonna *et al.*, 2012). DTW measures the similarity between two time series by allowing one of the series (query) to expanse or compress at each time point rather than point to point comparison such that the similarity is maximized. One of the key limitations of DTW is that, it is a global measure, i.e. DTW assumes that the two time series overlap on the edges and expanses one of the series to the length of other series to compute their similarity. However, in reality the gene expression is dynamic and the regulations are active only for a subset of time (Prelic *et al.*, 2006). This phenomenon is clearly observed in developmental and disease-prognosis networks (pathways) (Bar-Joseph, 2004). Thus in time-series gene-expression analysis, it is very important to detect local expression similarities to understand the underlying molecular dynamics in biological networks (Androulakis *et al.*, 2007). All these limitations of current methods advocates the need of new gene-network-inference methods, that are least influenced by the dynamic expression delays and the number of time points.

Here we present a new network inference method based on the gapped local alignment of gene-expression profiles. Gapped local alignment was originally used to align two nucleotide or protein sequences and to find the best matched subsequences (Smith and Waterman, 1981). Gaps induced in the alignment signify the insertions and deletions, and alignment score reflects the similarity between the two aligned sequences. Here, we employ gapped local alignment to infer dynamic delay GRN. The rationale behind the algorithm is that, if a gene *X* is regulating gene *Y*, the expression pattern of the target gene *Y* is stimulated by the expression pattern of its regulator *X*, i.e. they share similar expression pattern with varying expression delay. A regulatory relationship is often a local phenomenon due to the dynamic nature of gene regulation (e.g. different co-factor binding). Figure 1b and c shows the example local expression correlations (shaded). By identifying such common local patterns between a regulator and its target, we can reveal the relationships between them. Figure 1(d) shows positive regulation between the genes pal-1 and elt-1 in the C lineage of *Caenorhabditis elegans* (Murray *et al.*, 2012). It can be observed that the delay is not constant throughout the expression, i.e. $D1 \neq D2$ (Fig. 1d). Insertion of gaps in the alignment will incorporate dummy time points that can account for the variability in delay. Instead of mapping the expression values one to one, we stretch the expression values of one gene over other by inserting gaps, such that the similarity between the two expression patterns (regulator and its target) is maximized. A detailed description of the algorithm is given in the Methods section.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

2 METHODS

2.1 Gapped local alignment of expression profiles

Here, we employed dynamic programming approach to align two expression patterns with varying delay and number of time points. Consider the expression patterns of two genes A and B with x and y number of time points, respectively:

$$A = a_1, a_2, a_3, \cdots, a_x$$
$$B = b_1, b_2, b_3, \cdots, b_y.$$

First, a similarity matrix S of the order $x \times y$ is computed. We compute the similarity $s(i,j)$ as an exponential function of the distance between time points a_i and b_j.

$$s(i,j) = e^{-\alpha \times d(a_i, b_j)} \tag{1}$$

where α is the measure of steepness and $d(a_i,b_j)$ is the distance between time points a_i and b_j. A value of $\alpha = 1.7$ is used in this study (discussed in Supplementary Material). Euclidean distance is the most commonly used distance metric. However, it is heavily influenced by the magnitude of the difference between the data points. We are more interested in the trend of expression than the change in magnitude. Thus we capture the expression trend at a_i and b_j with respect to their neighboring data points as shown below (Supplementary Material):

$$t(a_i) = \left[(a_i - a_{i-1}) + (a_{i+1} - a_{i-1})/2\right]/2 \tag{2}$$

$$t(b_i) = \left[(b_i - b_{i-1}) + (b_{i+1} - b_{i-1})/2\right]/2 \tag{3}$$

The underlying rationale, i.e. a regulator and its target share a similar expression trend irrespective of huge variations in the magnitude. Distance between the two expression trends, $d(a_i, b_j)$ is computed as shown below:

$$d(a_i, b_j) = \left|t(a_i) - t(b_j)\right|. \tag{4}$$

Next, an alignment matrix M of the order $(x+1) \times (y+1)$ is computed. An extra row and column are added to allow gaps of any length in either of the expression patterns (Fig. 2a). Typically, an un-gaped alignment approximately requires n^2 computations, where n is the average length of the time series. If we consider gap insertions, the computational complexity increases exponentially. Thus, we adopted the dynamic programing approach to reduce the number of computations to compute the gapped alignment. The value of the (i, j)-th element in the alignment matrix M is computed from its adjacent cells in three possible ways (positions) (Fig. 2a), a diagonal position $(i-1)$ and $(j-1)$ with no gaps, or from $(i-1)$ and j with a gap inserted in series A, or from i and $(j-1)$ with a gap inserted in series B.

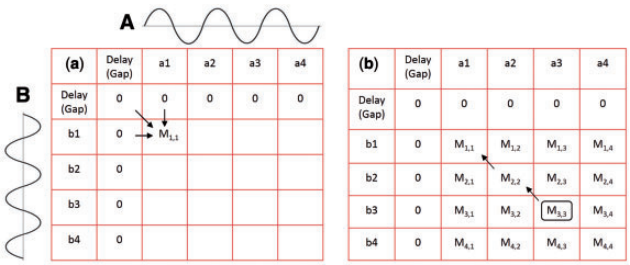

Fig. 2. (**a**) Alignment matrix M with an extra row and column to accommodate gaps and the three possible paths to compute an element $M_{1,1}$ and (**b**) shown in black are the recorded paths (tracebacks) used to compute the alignment

$$M_{ij} = \max \begin{Bmatrix} M_{i-1,j-1} + s'(i,j) \\ M_{i-1,j} - p \\ M_{i,j-1} - p \\ 0 \end{Bmatrix} \tag{5}$$

$M_{i,j}$ is the score for position i in series A and position j in series B, $s'(i, j)$ is the normalized similarity between time positions a_i and b_j (section 2.2), and p is the gap penalty. In the current study, a gap penalty of 0.3 is used (discussed in Supplementary Material). However, it is recommend that different gap penalties be used, based on the diversity of gene-expression patterns. At each step, the direction (path) of the highest score is recorded. Once the alignment matrix is completely filled, we compute the alignment by joining all the recorded paths starting from the maximum element in the matrix, as illustrated in Figure 2. Alignment score N is computed as:

$$N = \left[\frac{\mathrm{Max}(M)}{L}\right] \tag{6}$$

where, Max(M) is the maximum element in the alignment matrix M and L is the alignment length. Multiple alignments are possible if there is more than one path from the maximum element in the alignment matrix M.

2.2 Minimizing the effect of shift and scale

In the current context of gene-network inference, we are more interested in the expression trend of a gene than the actual change in its magnitude. Figure 3a shows the model-expression patterns of two genes with similar local trend, but high noise with different scale and shift makes it difficult to understand the trend similarity. Thus, the expression patterns are normalized prior to the alignment by subtracting the mean and dividing by the maximum:

$$T_i' = \frac{T_i - \overline{T}}{T_{\max}} \tag{7}$$

where, T_i is the i-th element, $\overline{T}$ is the mean and $T_{\max}$ is the maximum value of the time series T. This constrains the expression patterns between –1 and +1 with a unit variance and zero mean. The trend is more obvious after normalization as shown in Figure 3b. In a similarity matrix high similarity between any two time points might force the less similar neighboring points to align. To minimize this, we normalize the similarity matrix as follows:

$$s'(i,j) = \frac{s(i,j) - \bar{s}(i,j)}{m(i,j)} \tag{8}$$

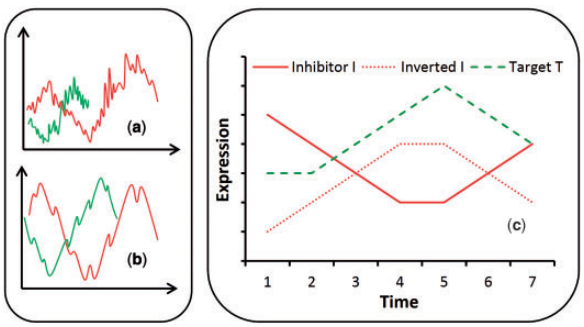

Fig. 3. Normalization and inversion of gene-expression patterns. (**a**) Model raw-expression patterns. (**b**) Normalized expression patterns. (**c**) Model inhibitor–target relationship

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

where, $s(i, j)$ is the (i, j)-th element in the similarity matrix, $\bar{s}(i, j)$ is the average and $m(i, j)$ is the maximum of the i-th row and j-th column of the similarity matrix S. By doing so, influence of the high similarity between the time points i and j is restricted to i-th row and j-th column.

2.3 Aligning inhibitors to targets

The expression patterns of inhibitor and its target are inverted on the time axis with respect to each other as shown in Figure 3c. To comprehend such relationships we flip the expression pattern of inhibitor by an inverse operation before aligning it to its target (shown in green, Fig. 3c). For every gene pair, we perform both direct and inverted alignments and the mode of regulation is inferred based on the high-scoring significant alignment.

2.4 Inferring regulator–target relationships

We infer a regulatory relationship between any two genes, if they share a common expression trend. As gene regulations are dynamic and vary with time, the observed correlations can be local with varying delay. From Figure 1c, we can observe a clear mutual trend between the gene pal-1 (regulator) and its target elt-1 (whose regulatory relationship is well established). Such common patterns can be identified by high (gapped local) alignment scores. The direction of regulation is inferred based on the onset time of gene expressions i.e., the gene with an early onset time is predicted as regulator and vice versa. In Figure 1c pal-1 is expressed first, and hence is predicted as the regulator of elt-1.

2.5 Significance of the alignment

Simple alignment score is not a clear indicator of a significant alignment as the expression pattern of a gene can be randomly aligned to the expression patterns of other genes. Conventionally, a P-value is calculated from the distribution of alignment scores to access the significance of an alignment (Altschul *et al.*, 1997; Pearson, 1996). The distribution of gapped-local-alignment scores follows an extreme value distribution (Bailey and Gribskov, 2002). The probability density function for the generalized extreme value distribution with location parameter μ, scale parameter σ and shape parameter $k \neq 0$ is of the form:

$$f(X|k,\mu,\sigma) = \left(\frac{1}{\sigma}\right)\exp\left(-\left(1+k\frac{(x-\mu)}{\sigma}\right)^{-1/k}\right)\left(1+k\frac{(x-\mu)}{\sigma}\right)^{-1(-1/k)} \quad (9)$$

To calculate the P-value, 100 000 random background alignment scores are computed by shuffling the normalized expression levels at random points (Li *et al.*, 2010). The parameters μ, σ and k are estimated based on these 100 000 random alignment scores. A P-value for a given alignment score is thus computed based on the estimated parameters. The P-value is expected to be as small as possible for a significant alignment. The R package 'evir' for extreme value distribution to estimate the parameters and to compute the P-value is used.

2.6 Multiple regulators

We use the window or the interval of regulation to predict collective gene regulators i.e. multiple regulators regulating their target at the same time. If a gene is collectively regulated by more than one gene at a time, they share a common window of regulation (aligned expression patterns) irrespective of their magnitudes. Figure 4a shows model-expression patterns of genes a and c collectively regulating b, in the same window/interval (w1 = w2). All the regulators that share a common regulatory window with respect to a target gene are can be inferred to be its collective regulators. Alternatively, a gene can also be regulated by multiple regulators at different time points (windows) as shown in Figure 4b. These relationships are relatively easier to infer as they have distinct regulatory windows (w1 ≠ w2). Since DDGni (Dynamic delay gene-network inference) builds a global network, at this level it is difficult to distinguish the regulators acting at different time points. However, DDGni also provides a separate complete alignment file with alignment coordinates. The order of multiple regulations can be inferred based on the order of aligned coordinates i.e. earlier the alignment earlier is its regulation. This facilitates a better understanding of the underlying network dynamics.

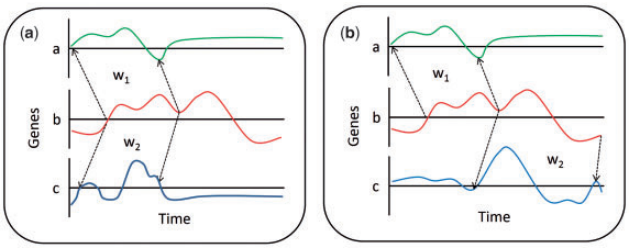

Fig. 4. Illustration of multiple regulatory relationships of a target gene. (**a**) Genes a and c collectively regulating b, in the same interval/window (w1 = w2). (**b**) Genes a and c regulating b at different intervals/windows (w1 ≠ w2)

3 RESULTS

The merit of the current method DDGni is to infer the dynamic time delayed GRN. Conventionally, multifactorial DREAM data is used to evaluate the performance of various GRN-inference methods. As the DREAM data is a steady-state data and does not include any time delay in the gene-expression patterns, it is thus rendered incompatible to illustrate the merits of DDGni on this dataset. Therefore, it was considered imperative to simulate data that incorporates time delay in the gene expressions in order to highlight the application of DDGni in handling dynamic time delay. Furthermore, to draw parallels with the already existing methods, TimeDelay-ARACNE (Zoppoli *et al.*, 2010), an information theoretic method, is also evaluated on the same simulated data. The results support the potential of DDGni in inferring time delayed GRNs (section 3.1).

Furthermore to demonstrate the applicability on steady-state data, DDGni is also evaluated on multifactorial DREAM 4 data (with no time delay), against three more prominent network inference methods: Maximal information coefficient (MIC) (Reshef *et al.*, 2011), ARACNE (Margolin *et al.*, 2006) and GENIE3 (Huynh-Thu *et al.*, 2010). GeneNetWeaver (Schaffter *et al.*, 2011) is used to generate the DREAM 4 multifactorial data. The performance of all four methods is quite comparable as shown in Supplementary Table S1.

To assess the practical application of DDGni, we used the cell-cycle time-course data from Yeast (Spellman *et al.*, 1998) and real time embryonic gene-expression data from *C.elegans* (Murray *et al.*, 2012), as discussed in section 3.2 and 3.3 respectively. The performance of DDGni is evaluated against TD-ARACNE and DTW. The AUC (Baldi *et al.*, 2000) values show a substantially increased performance of DDGni as compared to TD-ARACNE and DTW.

3.1 Simulated gene-expression dataset

We generated synthetic gene networks with different topologies, sizes and most importantly varying time delay as follows.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

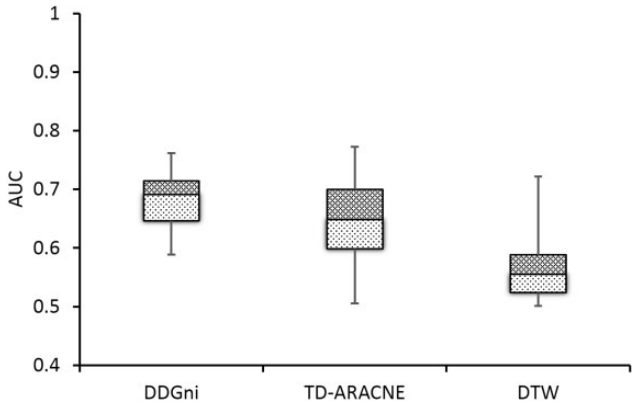

Fig. 5. Performance of DDGni, TD-ARACNE and DTW on the simulated gene-expression data with dynamic delay

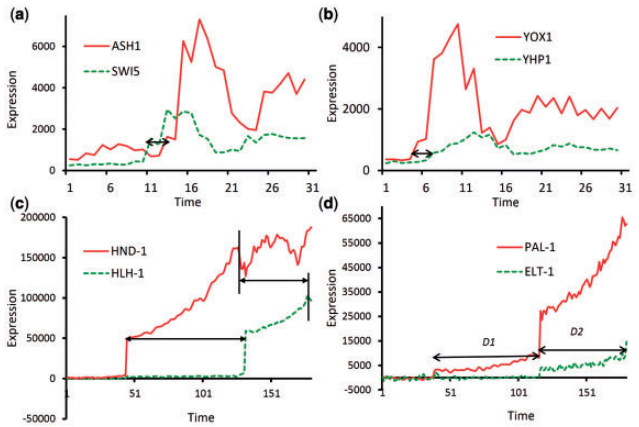

Fig. 6. Time delayed regulatory relationships between (**a**) SWI5:ASH-1, (**b**) YOX1:YHP1, (**c**) HND-1:HLH-1 and (**d**) PAL-1:ELT-1. Delays are marked by double-headed arrows

- A directed random acyclic graph is generated; each node here corresponds to a gene.
- If the number of genes with more than one regulator is <20%, we re-generate the network (Zoppoli *et al.*, 2010).
- Each gene G is initialized (100 time points) with numbers in [0, 1] following uniform random distribution.
- For each target gene T, the expression values are computed as a function of its regulator R:

$$T_{[i]} = \beta_i R_{[d+i]} + e_i \tag{10}$$

$$d = d_{\mathrm{i}}\mathrm{n} - d_{\mathrm{d}} \tag{11}$$

  where T is the target gene, R is the regulator, β_i is the uniformly distributed regulatory coefficient, e_i is the noise, d is the dynamic delay, d_{in} is the initial delay and d_{d} is the dynamic delay factor responsible for increase or decrease in the initial delay.
- Dynamic delay (d) is updated at every time point (i) according to the dynamic delay factor (d_{d}). The value of d_{d} is based on a random number n [0, 1], $d_{\mathrm{d}} = -1$ if $(n < x)$, 0 if $(x \leq n \leq y)$ and −1 if $(y < n)$. To minimize the large fluctuations in delay we used $x = 0.1$ and $y = 0.9$.

DDGni, TD-ARACNE and DTW are evaluated on 100 synthetic networks with dynamic delay (see Supplementary Material). The current problem boils down to a simple binary classification problem, i.e. to classify a gene pair as interacting (regulatory relationship) or non-interacting (no regulatory relationship). Conventionally, area under the receiver-operating characteristic (ROC) curve is used to evaluate the performance of binary classifiers with respect to its discrimination threshold (Madhamshettiwar *et al.*, 2012). ROC is a function of true positive rate (tpr) and false positive rate (fpr) (Baldi *et al.*, 2000). Alignment score is the discrimination factor here. The advantage of using AUC is that, we need not optimize a discrimination threshold. Figure 5 shows the range of AUC values of DDGni, TD-ARACNE and DTW, observed over the 100 simulated networks. The AUC values of DDGni are higher when compared to the other two (Fig. 5 and Table 1). This suggests the ability of DDGni to handle the dynamic delay embedded in the high-temporal gene-expression profiles. Performance evaluation at different noise levels suggests the robustness of DDGni (discussed in Supplementary Material).

Table 1. Performance (AUCs) of DDGni, TD-ARACNE and DTW on the three datasets

Methods	Simulated	Yeast	*C.elegans*
DDGni	0.68	0.74	0.60
TD-ARACNE	0.64	0.61	0.54
DTW	0.56	0.63	0.55

3.2 Cell cycle time-course data from Yeast

Next, we evaluated our method on a well-established network of eight transcriptional factors in Yeast. The cell cycle time-course gene-expression data is downloaded from GEO (GSE8799). The dataset consist of two replicates with 15 time points each. We merged both the replicates to get 30 time points. We selected eight TFs (YOX1, STB1, HCM1, WHI5, YHP1, ACE2, SWI5 and ASH1) that are extensively studied (Orlando *et al.*, 2008). The regulatory relationships among these eight TFs are obtained from literature, YEASTRACT (Abdulrehman *et al.*, 2011) and STRING (Szklarczyk *et al.*, 2011). Figure 6 shows the time delayed regulatory relationship between SWI5:ASH-1 (Fig. 6a) and YOX1:YHP1 (Fig. 6b). The AUC values reported in Table 1 show a substantial increase in the performance of DDGni when compared to other methods.

3.3 Embryonic gene-expression data from *Caenorhabditis elegans*

Recent studies have measured the gene expression in *C.elegans* embryo using live cell imaging techniques (Murray *et al.*, 2006). Diverse cell lineages and tissue types are the consequence of different cell fates and thus help comprehend underlying molecular

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

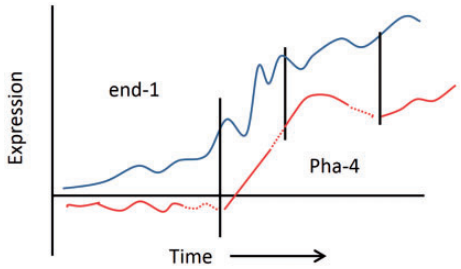

Fig. 7. Splined interpolation of the expression values to ensure equal cell-cycle lengths

mechanisms. To understand this, Bao *et al.* (2006) and Murray *et al.* (2008) have developed methods to quantitatively measure the gene expressions of every cell with 1-min interval during embryogenesis. Using these protocols Murray *et al.* (2012) have measured the expression values of 127 genes, mostly transcription factors with ~1 min temporal and single cell resolution in *C.elegans* embryo. The resulting expression patterns from these cell lineages are rich in information, including tissue types, cellular physical positions and cell division and symmetries. Inferring regulatory networks from such data will help us understand various regulatory mechanisms involved in tissue differentiation and embryonic development. Figure 6 shows the time delayed regulatory relationship between HND-1:HLH-1 (Fig. 6c) and PAL-1:ELT-1 (Fig. 6d) in the C lineage of *C.elegans*.

3.3.1 Interpolation of gene expression values The expression values of 127 genes are measured independently (one reporter gene per embryo) by quantifying the fluorescent reporter expression driven by their promoter sequences. Ideally, a cell should have the same cell-cycle length across all experiments. However, due to specific experimental conditions the cell-cycle lengths might vary across different embryos. This may be also due to non-uniform and irregular sampling that are common while tracking embryonic development (White *et al.*, 1999).

Figure 7 shows the differences in the cell division times for two genes (experiments) end-1 and pha-4 in the E lineage of *C.elegans* (cell division events are marked by a solid black line). It can be observed that pha-4 cell cycle lengths are shorter when compared to that of end-1. These differences should be handled before proceeding to their alignment. Interpolation technique is used to fit the pha-4 curve to end-1. B-splines are already used successfully on similar gene-expression data (Bar-Joseph *et al.*, 2003). The red dotted regions in Figure 7 are the interpolations of pha-4 expression.

3.3.2 Benchmark data To evaluate our method on real time embryonic gene-expression data, we manually curated a benchmark dataset from the available literature for which the embryonic gene expression data is available (Supplementary Table S5). The AUC values reported in Table 1 suggest an enhanced performance of DDGni as compared to TD-ARACNE and DTW.

The AUC of ROC curve can suggest the overall performance of a program without needing to consider the specific cutoffs.

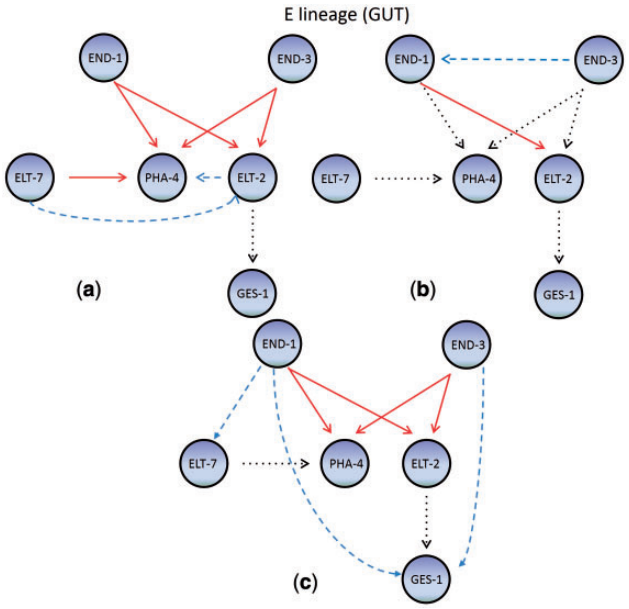

Fig. 8. Networks inferred from (**a**) DDGni, (**b**) TD-ARACNE and (**c**) DTW. Solid lines, true positives; dotted lines, false negatives; dashed lines, false positives

However, TD-ARACNE is a binary classifier, i.e. it outputs 1 or 0 corresponding to the presence or absence of a regulatory edge, respectively. Thus, we also evaluated the performance in terms of *F*-score (Powers, 2011) (Supplementary Table S4). Computational (time) complexity and running time of respective methods are reported in Supplementary Table S7. The overall tradeoff between AUC and runtime of DDGni is quite acceptable.

3.3.3 Evaluation on a well-established network For a better illustration of the merits of DDGni, we demonstrate the performance of our method in comparison to TD-ARACNE on a well-established network in *C.elegans*. Expression of end-1, end-3, elt-7, pha-4, ges-1 and elt-2 genes are specific to the E lineage of *C.elegans* and are important for the gut development. Figure 8 shows the networks inferred by DDGni (Fig. 8a), TD-ARACNE (Fig. 8b) and DTW (Fig. 8c); solid lines are the true relationships that are also predicted by the respective methods, dotted lines are the true relationships that are not predicted and dashed lines are the predictions that are not true (not reported in literature). From Figure 8, we observe that predictions by DDGni overlap more with the true relationships when compared to TD-ARACNE and DTW.

4 CONCLUSION

Although, several methods are proposed to infer regulatory networks from temporal data, their performance is not satisfactory, especially with the dynamic delay associated to the gene regulation. High complexity and limitations of the existing methods in handling varying time delay, advocates the need of effective gene network inference methods that are least influenced by expression delays and number of time points. In this study, we

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

proposed a simple and elegant network inference method based on gapped local alignment of gene expression profiles. By identifying common local patterns between a regulator and its target, we can reveal the regulatory relationship between them. In short, we span the expression values (time points) of the target gene over the expression values of its potential regulator by inserting gaps such that the similarity between the two expression patterns is maximized. The novelty of our method is the use of 'gaps' to handle the dynamic delay in gene regulation and uniformly sampled time points which is quite common in long time series such as cell-lineage data. The order of multiple regulations can be inferred based on the order of aligned coordinates, i.e. earlier the alignment earlier is its regulation. The proposed method is computationally less complex and exercise dynamic programming. We evaluated our performance against prominent network inference methods like TD-ARACNE, DTW, MIC, ARACNE and GENIE3. The AUC values for both real time and simulated time series gene-expression data evince an improved performance by our proposed method in handling the dynamic delay during transcriptional regulation and the evaluation on steady state DERAM4 data suggests the on-par performance of DDGni with other prominent methods. DDGni is highly suitable for real temporal data with high sampling frequency where delay dynamics is obvious. It is also applicable to short time-series data as suggested by its performance on the yeast cell-cycle data. In addition to the above, its on-par performance on static data advocates a more general applicability. However, as any other pure expression driven method it suffers from spurious relationships, i.e. identified correlations (delayed-similarity) do not represent true causal relationships. This problem can be solved by using ChIP-Seq binding data (Qin *et al.*, 2011). However, in this article only expression data is used for network construction.

Funding: Research Grants Council, Hong Kong SAR, China (grant number 781511M); National Natural Science Foundation of China, China (grant number 91229105).

Conflict of interest: None declared.

REFERENCES

Aach,J. and Church,G.M. (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics*, **17**, 495–508.

Abdulrehman,D. *et al.* (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res.*, **39**, D136–D140.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Androulakis,I.P., Yang,E. and Almon,R.R. (2007) Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annu. Rev. Biomed. Eng.*, **9**, 205–228.

Bailey,T.L. and Gribskov,M. (2002) Estimating and evaluating the statistics of gapped local-alignment scores. *J. Comput. Biol. A J. Comput. Mol. Cell Biol.*, **9**, 575–593.

Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Bao,Z. *et al.* (2006) Automated cell lineage tracing in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA*, **103**, 2707–2712.

Bar-Joseph,Z. (2004) Analyzing time series gene expression data. *Bioinformatics*, **20**, 2493–2503.

Bar-Joseph,Z. *et al.* (2003) Continuous representations of time-series gene expression data. *J. Comput. Biol. A J. Comput. Mol. Cell Biol.*, **10**, 341–356.

Blake,W.J. *et al.* (2003) Noise in eukaryotic gene expression. *Nature*, **422**, 633–637.

Bratsun,D. *et al.* (2005) Delay-induced stochastic oscillations in gene regulation. *Proc. Natl Acad. Sci. USA*, **102**, 14593–14598.

Davidson,E. and Levine,M. (2005) Gene regulatory networks. *Proc. Natl Acad. Sci. USA*, **102**, 4935–4935.

Huang,T. *et al.* (2010) Using GeneReg to construct time delay gene regulatory networks. *BMC Res. Notes*, **3**, 142.

Huynh-Thu,V.A. *et al.* (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.

Josic,K. *et al.* (2011) Stochastic delay accelerates signaling in gene networks. *PLoS Comput. Biol.*, **7**, e1002264.

Lee,C.-P., Leu,Y. and Yang,W.-N. (2012) Constructing gene regulatory networks from microarray data using GA/PSO with DTW. *Applied Soft Comput.*, **12**, 1115–1124.

Li,M.J., Sham,P.C. and Wang,J.W. (2010) FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution. *Bioinformatics*, **26**, 2897–2899.

Madhamshettiwar,P.B. *et al.* (2012) Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Med.*, **4**, 41.

Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. USA*, **107**, 6286–6291.

Margolin,A.A. *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform.*, **7** (**Suppl. 1**), S7.

Murray,J.I. *et al.* (2006) The lineaging of fluorescently-labeled *Caenorhabditis elegans* embryos with StarryNite and AceTree. *Nat. Protoc.*, **1**, 1468–1476.

Murray,J.I. *et al.* (2008) Automated analysis of embryonic gene expression with cellular resolution in *C.elegans*. *Nature Methods*, **5**, 703–709.

Murray,J.I. *et al.* (2012) Multidimensional regulation of gene expression in the *Caenorhabditis elegans* embryo. *Genome Res.*, **22**, 1282–1294.

Orlando,D.A. *et al.* (2008) Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature*, **453**, 944–947.

Pearson,W.R. (1996) Effective protein sequence comparison. *Methods Enzymol.*, **266**, 227–258.

Powers,D.M.W. (2011) Evaluation: from precision, recall and *F*-factor to ROC, informedness, Markedness & Correlation. *J. Machine Learn. Technol.*, **2**, 37–63.

Prelic,A. *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Qin,J. *et al.* (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Res.*, **39**, W430–W436.

Reshef,D.N. *et al.* (2011) Detecting novel associations in large datasets. *Science*, **334**, 1518–1524.

Rhudy,M. *et al.* (2010) Microphone array analysis methods using cross-correlations. *Imece2009, Vol 15: Sound, Vibration and Design*, 281–288.

Riccadonna,S. *et al.* (2012) DTW-MIC coexpression networks from time-course data. arXiv:1210.3149 [q-bio.MN].

Schaffter,T. *et al.* (2011) GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**, 2263–2270.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stolovitzky,G. *et al.* (2009) Lessons from the DREAM2 Challenges. *Ann. New York Acad. Sci.*, **1158**, 159–195.

Szklarczyk,D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.

White,K.P. *et al.* (1999) Microarray analysis of Drosophila development during metamorphosis. *Science*, **286**, 2179–2184.

Zhu,R. *et al.* (2007) Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models. *J. Theor. Biol.*, **246**, 725–745.

Zoppoli,P. *et al.* (2010) TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinform.*, **11**, 154.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016