BIOINFORMATICS

Vol. 18 no. 1 2002
Pages 61–66

Extracting transcriptional events from temporal gene expression patterns during Dictyostelium *development*

R. Šášik[1,], N. Iranfar[2], T. Hwa[1] and W. F. Loomis[2]*

[1]Department of Physics and [2]Division of Biology, University of California at San Diego, La Jolla, CA 92093, USA

Received on April 19, 2001; revised on July 16, 2001; accepted on August 20, 2001

ABSTRACT

Motivation: The DNA microarray technology can generate a large amount of data describing the time-course of gene expression. These data, when properly interpreted, can yield a great deal of information concerning differential gene expression during development. Much current effort in bioinformatics has been devoted to the analysis of gene expression data, usually via some 'clustering analysis' on the raw data in some abstract high dimensional space. Here, we describe a method where we first 'process' the raw time-course data using a simple biologically based kinetic model of gene expression. This allows us to reduce the vast data to a few vital attributes characterizing each expression profile, e.g. the times of the onset and cessation of the expression of the developmentally regulated genes. These vital attributes can then be trivially clustered by visual inspection to reveal biologically significant effects.

Results: We have applied this approach to microarray expression data from samples isolated every 2 h throughout the 24 h developmental program of *Dictyostelium discoideum*. mRNA accumulation patterns for 50 developmental genes were found to fit the kinetic model with a p-value of 0.05 or better. Transcription of these genes appears to be initiated in bursts at well-defined periods during development, in a manner suggestive of a dependent sequence. This approach can be applied to analyses of other temporal gene expression patterns, including those of the cell cycle.

Contact: sasik@physics.ucsd.edu

Supplementary information: Intensity ratios for all genes in this study are available at http://www.biology.ucsd.edu/loomis-cgi/microarray/index.html.

INTRODUCTION

Dictyostelium discoideum is a social amoeba that lives in the soil, feeding on bacteria (Loomis, 1975). Upon starvation, a 24 h long developmental program ensues, leading eventually to the formation of a differentiated multicellular organism in which 80% of the original cells have become resistant spores and the remaining 20% have become stalk cells that physically support the spores. During development, *D.discoideum* shares many of the physiological functions seen in mammalian cells such as directed amoeboid movement, cell–cell adhesion, tissue differentiation, proportioning, and sorting (Loomis, 1996).

Immediately following the initiation of development, growth stops and the previously solitary cells express genes whose products permit them to interact and form aggregates containing up to 10^5 cells. Within 8 h, they can be seen streaming into centers where they form hemispherical mounds. During the following 16 h, these multicellular structures go through a stereotyped series of morphological changes leading to the formation of fingers at 12 h, and eventually fruiting bodies; see the cartoon of development depicted in Figure 1. Conditions have been found that result in the *synchronous* development of thousands of such fruiting bodies starting from 10^9 genetically identical cells; these have led to high resolution biochemical and molecular studies over the last three decades (Loomis, 1975, 1996). Specific mRNAs and proteins have been found to accumulate at various stages of development using Northern analyses and 2-D gels (Chung *et al.*, 1981; Loomis, 1985). Once the finger stage is reached, few new genes are expressed until the beginning of culmination at 17 h (Chung *et al.*, 1981; Loomis, 1985). Due to technical limitations and the rather qualitative nature of these experimental techniques, the early studies focused on only a few genes at a time, and the order of expression of the genes could only be crudely inferred.

We have used microarrays of nearly 700 selected genes to define their relative expressions at 2 h intervals throughout development. For the *temporal* expression patterns at hand, we use a natural approach based on the first-order kinetics of mRNA abundance in the cell.

*To whom correspondence should be addressed.

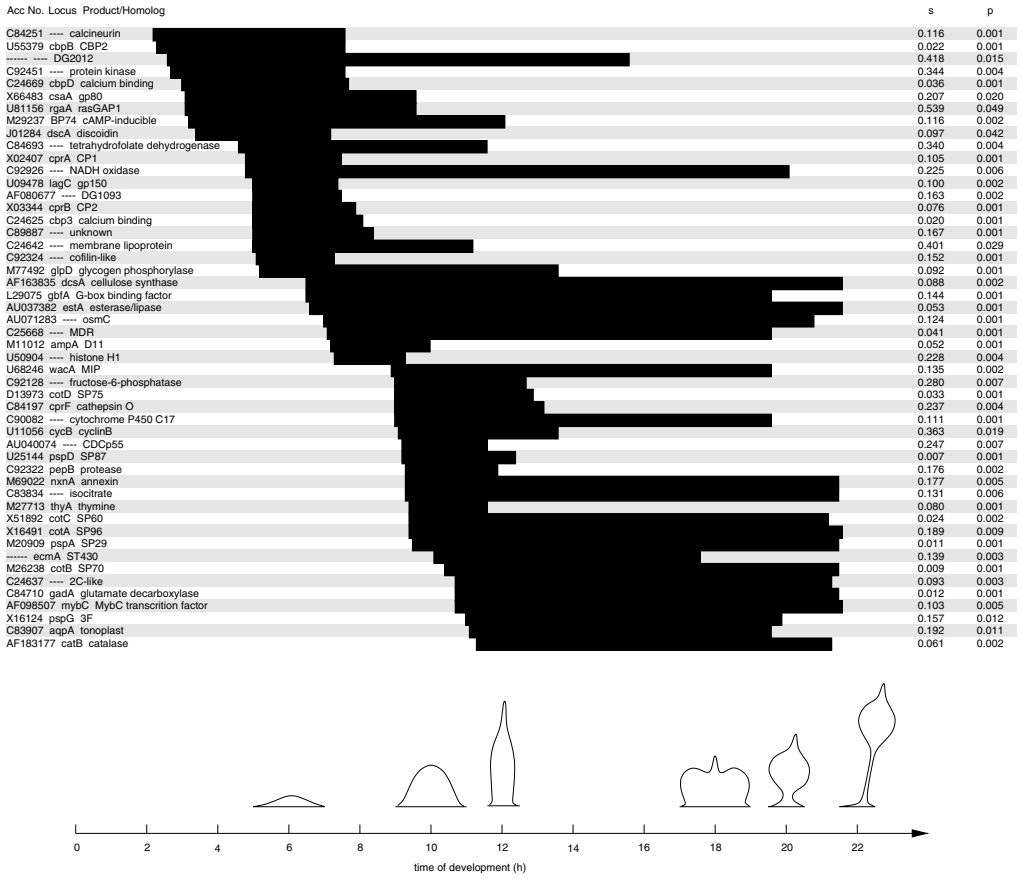

Fig. 1. Transcription regulation of 50 developmental genes of *D.discoideum*, with corresponding *s*-statistics and *p*-values of the fit to the kinetic model. Transcription of a gene is initiated at the beginning of the black band to the right of the gene descriptor, and terminates at the end of the black band. The grey-and-white stripes are a guide to the eye.

It allows us to extract the time at which each gene is first expressed, the half-life of its mRNA, and the time at which its expression ceases. The results obtained validate and quantify the above-mentioned qualitative observations made during the past three decades, and point towards future experiments with null mutant strains that will disclose the place and function of individual genes in the genetic network that directs the developmental program of *D.discoideum*.

METHODS

Gene probes were microarrayed robotically on glass slides (Iranfar *et al.*, 2001). The entire genome of *D.discoideum* is estimated to contain 8000–10 000 genes, and is in the process of being sequenced. (At the present time about 6000 genes have been identified.) The genes used in this study included previously characterized developmentally regulated genes as well as genes encoding proteins with significant similarity to proteins characterized in other organisms but not previously encountered in *D.discoideum*.

A large number ($\sim 10^9$) of identical vegetative cells of wild-type strain NC4 were induced to initiate development synchronously by the removal of nutrients and spreading on a buffered surface. Standard methods were used to isolate total RNA from $\sim 10^8$ cells at each time point and to subsequently collect poly-A^+ mRNA. The RNA preparations were used as templates for reverse transcriptase to generate cDNA copies of each mRNA. In order to determine relative temporal changes in specific gene expression free of slide- or probe-specific properties, a reference mixture of mRNAs collected from several experimental time points was copied into cDNA in the presence of dCTP-Cy5 (red) fluorescent dye, while the RNAs collected at each time point were copied in the presence of dCTP-Cy3 (green) fluorescent dye. Microarrays were hybridized for 18 h with a mixture of approximately equal amounts of Cy3 and Cy5 labelled DNA (Lockhart *et al.*, 1996; De Risi *et al.*, 1997). The fluorescent intensities from each spot was measured and processed.

ALGORITHM

The experiment produces a temporal sequence of expression levels $E_i(t)$ for each gene i, where t denotes the time of measurement. As a matter of course, an analyst might then proceed to cluster the expression profiles using one of the existing clustering algorithms (Jain and Dubes, 1988). The input to any such clustering analysis is a set of distances, each of which defines the degree of similarity between a pair of expression profiles. The distance measure is often taken to be a simple mathematical function of the expression profiles, such as the 'Euclidean distance' $D_{ij} = \left[\sum_t (E_i(t) - E_j(t))^2\right]^{1/2}$. While such a distance measure might be appropriate in situations where no obvious relations exist among the experiments, it is not the optimal one to use in analyzing *temporal* expression profiles where the experiments are causally linked. For example, the distance D_{ij} defined above is invariant with respect to permutations of the temporal sequence, even though the permuted sequences have no physical or biological meaning. In the ensuing analysis, we will exploit the causality of the temporal expression profiles to extract the underlying biologically relevant transcriptional events. The clustering analysis then becomes a trivial exercise as we shall see below.

Developmental genes are typically not being transcribed in vegetative feeding cells, but become expressed at various stages after initiation. As is well established (see, e.g. Singer and Penman, 1973) we model the abundance $A_i(t)$ of any particular gene transcript i at time t by a first-order differential equation

$$\frac{\mathrm{d}A_i(t)}{\mathrm{d}t} = S_i(t) - \gamma_i A_i(t), \tag{1}$$

where $S_i(t)$ is a gene-specific transcription regulation term, and γ_i is a gene-specific decay rate. In terms of the half-life τ_i of the transcript, $\gamma_i \equiv \ln(2)/\tau_i$[†]. The regulation term $S_i(t)$ is in general a complex function of all biologically relevant factors, including the abundance of transcription factors, enzymes, energetic resources of the cell, as well as external physical and chemical stimuli. In general, in order to describe the biochemical processes in the cell in their entirety, one has to deal with a large set of differential equations of this kind, one for each chemical involved. Approaches following this line of thinking have been developed (Chen *et al.*, 1999), but since most of the rate constants are not known quantitatively and many of the relevant interactions are not known even qualitatively, the utility of these methods is mostly academic. By contrast, in our approach the implicit time-dependence of all these factors appears as an explicit time-dependence in $S_i(t)$. We assume that the transcription regulation is a *sharp* function of time, e.g. no transcription occurs until the concentration of the necessary transcription factors and signaling molecules exceeds some threshold[‡], and model this process as a sequence of simple on and off events:

$$S_i(t) = \begin{cases} S_i & \text{if } t_1^i < t < t_2^i, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Thus in this model, transcription is initiated at time t_1^i and proceeds at a constant rate S_i until time t_2^i when it is terminated.

The formal solution of (1) and (2) is

$$A_i(t) = \begin{cases} 0, & t \leqslant t_1^i, \\ (S_i/\gamma_i)(1 - \mathrm{e}^{-\gamma_i(t-t_1^i)}), & t_1^i < t \leqslant t_2^i, \\ (S_i/\gamma_i)(1 - \mathrm{e}^{-\gamma_i(t_2^i - t_1^i)})\,\mathrm{e}^{-\gamma_i(t-t_2^i)}, & t_2^i < t. \end{cases} \tag{3}$$

In order to extract the transcription events t_1^i and t_2^i, we would like to fit the actual mRNA abundance levels $E_i(t)$ for each gene i with the solution (3), by minimizing the sum of squares $\sum_t [A_i(t) - E_i(t)]^2$ in the space of parameters t_1^i, t_2^i, S_i and γ_i. This can be done by means of any standard numerical package. However, $E_i(t)$ cannot be directly measured in an microarray experiment because of the different hybridization efficiency of the different DNA molecules. Instead, we follow the common procedure and approximate $E_i(t)$ by the ratio of background-adjusted[§] fluorescence intensities in the green and red channels taken from the target location on the microarray corresponding to gene i. As every expression pattern determined in this way contains information on mRNA abundance *relative* to the reference mixture (which, as explained above, is approximately an average of the abundance level at different times), $E_i(t)$ is dimensionless and can be scaled by an arbitrary multiplicative constant. We choose a scaling such that the mean expression level of every gene throughout development is unity. Quantities t_1^i, t_2^i and γ_i are not affected by the scaling.

We note in passing that the present approach can be easily extended to *periodic* gene expression data as well (such as the cell-cycle data). In that case the regulation term $S_i(t)$ is periodic in time and one looks for the periodic solution of (1).

[†] In principle, the half-lives themselves can be time-dependent. But since there is currently no sufficient data to extract this dependence, we will take them to be time-independent in this study.

[‡] This occurs because the bindings of transcription factors and the RNA polymerase are typically very sharp (i.e. Fermi) function of the concentration.

[§] Although every effort is made to use the same amount of poly-A$^+$ mRNA for the labeling reaction, the amount of dye incorporated into the cDNA probe varies across the time points, so in order to maintain compatibility among microarrays, we adopt a normalizing procedure in which the green fluorescent intensity of each target on a microarray is scaled by an overall multiplicative constant—common to that microarray—so that the *net* fluorescence intensities in the two channels on the same microarray are equal.

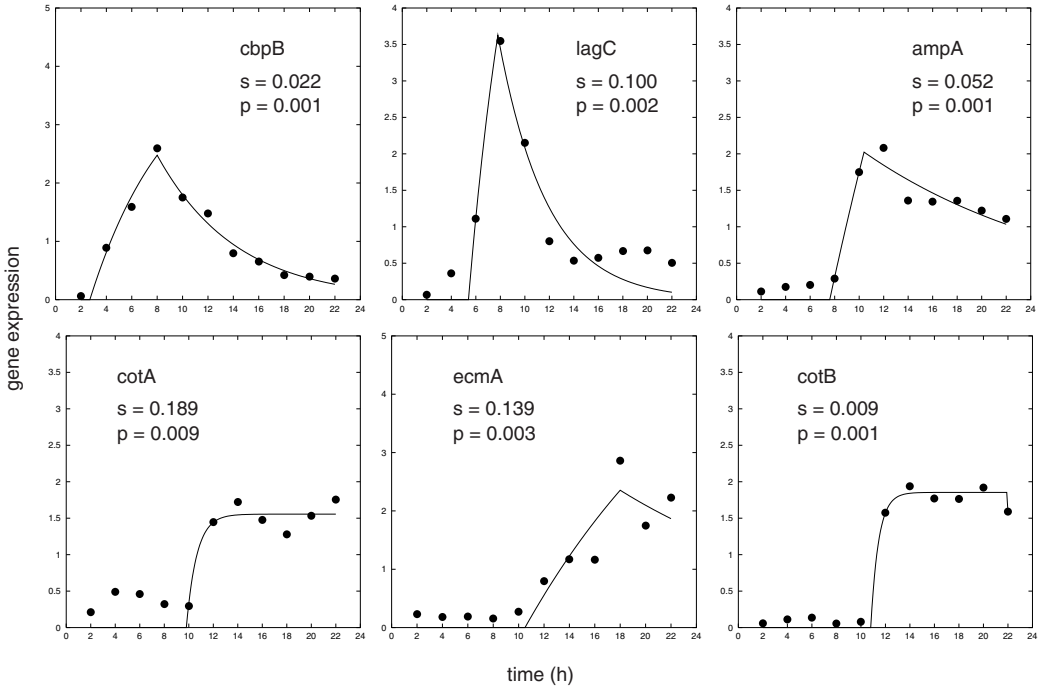

Fig. 2. Experimental expression patterns (points) with corresponding kinetic model fits (lines), with corresponding s-statistics and p-values, for six representative genes: *cbpB, lagC, ampA, cotA, ecmA,* and *cotB*.

The goodness of fit is measured by the statistic (omitting the indices)

$$s = \frac{\sum_t [E(t) - A(t)]^2}{\sum_t [E(t) - 1]^2}. \qquad (4)$$

The denominator in the above definition measures the amount of temporal variation found experimentally for a particular transcript, hence s is a measure of how much of that variation can be accounted for by our simple kinetic model. In general terms, a small s-statistic indicates that the expression pattern can be well described by the kinetic model. However, in order to be able to assess the statistical significance (e.g. the p-value) of the s-statistic found for any particular expression pattern, we need to know the distribution of the s-statistics found in a null sample of expression patterns that do *not* conform to the kinetic model. We estimate this distribution by a resampling method, in which the temporal sequence of the experimental expression pattern is permuted a large number of times, and the best fit $A(t)$ and the s-statistic are calculated for each permutation. This is precisely the approach motivated at the beginning of this section, where we identified the key signature of a physically and biologically meaningful *temporal* expression profile. With this method of significance assessment, the p-value is then given by the fraction of permutations whose s-statistics are smaller or equal to that found for the original experimental expression pattern. There are a total of 11 measurements taken; hence there are $11! \approx 4 \times 10^7$ permutations of every expression pattern. We use a random sample of 1000 permutations for calculation of the p-values. These should be properly labeled as *unadjusted* p-values (Westfall and Young, 1993), for we are in fact carrying out many statistical tests in the same family and no adjustment for multiple testing is being done.

DISCUSSION

The times of onset and cessation of expression for 50 developmentally regulated genes are shown graphically in Figure 1. The unadjusted p-values were calculated for a subset of 59 genes (of the original 673) whose transcription onset occurred after 2 h of development[¶]. Only the genes whose p-value was less than 0.05 were included in this figure. The sum of the p-values for the 50 listed genes is 0.3, indicating that there may be one or two genes whose conformity to the kinetic model is spurious. The median half-life of these 50 gene transcripts was 3.95 h, in good agreement with Chung *et al.* (1981), where a characteristic lifetime of 4 h is quoted. In Figure 2, we show some typical time courses and the best fit to the solution of first-order kinetics (equation 3) along with

[¶] The mRNA levels from the majority of the 673 genes did not increase more than 2-fold at any stage of development and were not further analyzed.

their s-statistics and p-values. One can view the onset and cessation time shown in Figure 1 as a 'filtered' version of the original expression profiles.

Of the numbers obtained from the least-square fit, the time of transcription onset is the most robust value, while the values of half-life and time of cessation of expression are subject to more uncertainty[‖]. A histogram of the onset times (t_1) is shown in Figure 3. In addition to those listed in Figure 1, we included the late spore specific gene *spiA* in this analysis even though it has a large p-value (0.4) because of its very late onset; however, previous Northern analyses have shown that its mRNA is not present before 18 h of development and then accumulates during culmination (Richardson *et al.*, 1991). A 'clustering analysis' in one dimension is best done by visual inspection. It is apparent that there are at least four distinct clusters of genes expressed during the first 12 h of development. The time intervals (gaps) between these clusters are 1.1, 1.3 and 1.5 h long. Can we state with confidence that this clustering is statistically significant and that transcription initiation of groups of genes is coordinated by design? If we assume that the observed clustering of onset times is a result of random grouping of random events uniformly spread out in the time window 2–12 h, the probability that the three largest gaps have size 1 h or larger is $\sim 3 \times 10^{-4}$. We conclude therefore that the observed clustering of transcription initiation events is highly statistically significant.

Thus, gene transcription appears to occur in coordinated bursts at roughly 3, 5, 8 and 10 h after the initiation of development with no new genes expressed until culmination about 5 h later. This pattern is similar to that described for newly synthesized proteins using 2-D gels (Loomis, 1985). It raises the possibility that expression of genes in the early clusters is required for subsequent expression of genes in later clusters. There is some genetic evidence supporting such a dependent sequence of events during development of *D.discoideum* (Loomis *et al.*, 1976; Loomis, 1998). When the complete genome sequence is available, the current studies can be expanded to a genome-wide survey which will undoubtedly uncover other genes that fall into these clusters. The roles of specific genes can then be assessed by characterizing mutant strains generated by molecular genetic techniques.

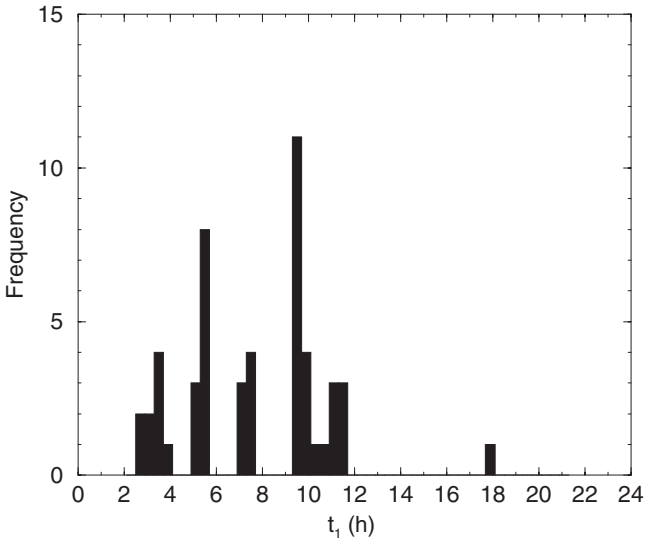

Fig. 3. Histogram of transcriptional onset times t_1 for 51 genes. The microarray data did not allow genes expressed before 2 h of development to be included in this analysis. Likewise, genes only expressed after 18 h were not sufficiently defined to be included except for *spiA*.

ACKNOWLEDGEMENTS

This research is supported by the Burroughs–Wellcome Fund through an Innovation Award in Functional Genomics to T.H. and a LJIS post-doctoral training fellowship to R.Š. T.H. further acknowledges the support of NSF through grant no. DMR-9971456 and BIO-0083704 (together with W.F.L.). N.I. and W.F.L. acknowledge the support of the NSF through grant no. 9728463, and the NIH through grant no. GM60447.

[‖] In some cases, there is more than one pair of values for the half-life and cessation time that produce nearly identical fits to the present data as measured by the sum-of-squares criterion.

REFERENCES

Chen,T., He,H.L. and Church,G.M. (1999) Modeling gene expression with differential equations. *Pac. Symp. Biocomput.*, **1999**, 29–40.

Chung,S., Landfear,S.M., Blumberg,D.D., Cohen,N.S. and Lodish,H.F. (1981) Synthesis and stability of developmentally regulated *Dictyostelium* mRNAs are affected by cell–cell contact and cAMP. *Cell*, **24**, 785–797.

De Risi,J. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Iranfar,N., Fuller,D., Sasik,R., Hwa,T., Laub,M. and Loomis,W.F. (2001) Expression patterns of cell-type specific genes in *Dictyostelium*. *Mol. Biol. Cell.*, in press.

Jain,A.K. and Dubes,R.C. (1988) *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ.

Lockhart,D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.*, **14**, 1675–1680.

Loomis,W.F. (1975) *Dictyostelium discoideum: A Developmental System*. Academic Press, New York.

Loomis,W.F. (1985) Regulation of cell-type-specific differentiation in *Dictyostelium*. *Cold Spring Harbor Symposia on Quantitative Biology*. pp. 769–777.

Loomis,W.F. (1996) Genetic networks that regulate development in *Dictyostelium* cells. *Microbiol. Rev.*, **60**, 135–150.

Loomis,W.F. (1998) Role of PKA in the timing of developmental events in *Dictyostelium* cells. *Microbiol. Mol. Biol. Rev.*, **62**, 684–694.

Loomis,W.F., White,S. and Dimond,R.L. (1976) A sequence of dependent stages in the development of *Dictyostelium discoideum*. *Dev. Biol.*, **53**, 171–177.

Richardson,D.L., Hong,C.B. and Loomis,W.F. (1991) A prespore gene, Dd31, expressed during culmination of *Dictyostelium discoideum*. *Dev. Biol.*, **144**, 269–280.

Singer,R.H. and Penman,S. (1973) Messenger RNA in HeLa cells: kinetics of formation and decay. *J. Mol. Biol.*, **78**, 321–334.

Westfall,P.H. and Young,S.S. (1993) *Resampling-based Multiple Testing*. Wiley, New York.