MATHEMATICAL MODELING OF ALTERNATIVE POLYADENYLATION IN THE HUMAN GENE, CSTF3

YIMING CHENG†, BIN TIAN‡, AND ROBERT M. MIURA§

Abstract. A large number of human genes have been found to contain more than one polyadenylation site (poly(A) site). These lead to multiple messenger ribonucleic acid (mRNA) transcripts, although some of them do not yield protein products. Like alternative initiation and alternative splicing, alternative polyadenylation (APA) contributes to the complexity of the overall pool of mRNA transcripts in human cells. However, little is known about how the cell utilizes different poly(A) sites in response to different cell conditions and developmental stages. Several protein factors have been identified as being necessary for in vitro cleavage and polyadenylation. If the gene of a polyadenylation factor has multiple poly(A) sites, producing variable protein products with different functions, then their APA can control the polyadenylation activity under different cell conditions. The dynamics of the gene expression levels among APA products so far has been studied by experimental approaches only. Here we propose a novel mathematical model consisting of differential equations with a time delay to study the dynamical behavior of the gene expression levels. Numerical simulations have been provided for some of the theoretical analyses of the equations. Our model suggests an oscillatory pattern of APA of CSTF3, indicating APA is employed to maintain homeostasis of its function, perhaps general polyadenylation activity as well, in the cell.

Key words. mathematical modeling, differential-delay equations, gene expression, alternative polyadenylation, cleavage and stimulatory factor 77

AMS subject classifications. 97M60, 92B05

DOI. 10.1137/120887266

1. Introduction. Pre-messenger ribonucleic acid cleavage and polyadenylation, also known as mRNA polyadenylation, is the cellular process that adds polyadenylation (poly(A)) tails to maturing mRNAs. The process of polyadenylation is composed of two tightly coupled steps [7]: an endonucleolytic cleavage at the poly(A) site and polymerization of an adenosine tail at the 3’ end of the cleaved RNA. Polyadenylation is directly linked to the termination of transcription [5]. Malfunction of polyadenylation has been implicated in several human diseases, such as immunodysregulation polyendocrinopathy enteropathy X-linked syndrome [2] and thrombophilia [13].

It has been reported that about 54% of human genes and 32% of mouse genes have multiple poly(A) sites [47]. Several protein factors have been identified to be necessary for in vitro cleavage and polyadenylation [23], including the cleavage and polyadenylation specificity factor (CPSF) [3, 14, 32], cleavage stimulation factor (CstF) [14, 45], cleavage factors I and II [38, 46], poly(A) polymerase (PAP) [46], and poly(A) binding protein [49]. CstF is composed of three subunits (50, 64, and 77 kDa). CstF-64

*Received by the editors August 7, 2012; accepted for publication (in revised form) June 17, 2013; published electronically September 10, 2013.

http://www.siam.org/journals/siap/73-5/88726.html

†Gene Center Munich, Department of Biochemistry, Ludwig-Maximilians-Universität München, Feodor-Lynen-Str. 25, 81377 Munich, Germany (cheng@genzentrum.lmu.de). This author’s work was supported by a Presidential Strategic Initiative Scholar Award from NJIT and Dr. Bin Tian’s laboratory from the University of Medicine and Dentistry of New Jersey (UMDNJ).

‡Department of Biochemistry and Molecular Biology, New Jersey Medical School, University of Medicine and Dentistry of New Jersey, Newark, NJ 07101 (btian@umdnj.edu). This author’s work was supported by The Foundation of UMDNJ.

§Department of Mathematical Sciences and Center for Applied Mathematics and Statistics, New Jersey Institute of Technology, Newark, NJ 07102 (miura@njit.edu). This author’s work was supported by NJIT (Department of Mathematical Sciences) and the National Science Foundation.
and CstF-77 are conserved across all eukaryotes. CstF-64 directly binds to U/GU-rich elements of transcripts containing a poly(A) signal [36], and CstF-77 has been shown to interact with several factors involved in cleavage and polyadenylation, such as CstF-50 [43], CstF-64 [21], CPSF-160 [33], and RNA polymerase II (RNAP II) C-terminus domain [29]. CstF-77 can also interact with itself [44].

The human gene, CSTF3 encoding CstF-77, has been found to contain an alternative polyadenylation site in the third intron [35] (see Figure 1.1). This poly(A) site is sometimes called an intronic poly(A) site. CstF-77.L refers to the “long” transcript generated by using the downstream poly(A) site and CstF-77.S refers to the “short” transcript generated by using the intronic poly(A) site. Surprisingly, the expression levels of the two transcripts show a very interesting pattern based on serial analysis of gene expression (SAGE) data as shown in Figure 1.2. It appears that CstF-77.L and CstF-77.S have opposite expression patterns: if the expression of the short form transcript increases, the long form expression decreases, and vice versa. However, the underlying regulation mechanism(s) are not clear, and they may be more complicated than observed. The long form protein is a key factor for polyadenylation. However, the function of the short form protein is still unknown.

In this study, we propose a mathematical model for the expression levels of CstF-77.L and CstF-77.S based on some assumptions. We incorporate a time delay in our equations to reflect translation of mRNA into protein. We derive a pair of differential-delay equations for the expression levels of the long and short form transcripts. Our approach will be to estimate the model parameters with hypothetical data due to the limited amount of reliable experimental data. Then we determine the consequences of this fully specified model problem. Bifurcation results and basins of attraction have been found by numerical simulation, and some theoretical analysis has been applied to understand the dynamical system.
2. Proposed model. The short form transcript of CSTF3 in human and mouse was first described in [35] and its function has yet to be understood. This transcript is generated by an intronic polyadenylation site (see Figure 1.1), which has been studied widely [4, 26, 48]. However, regulation among the transcripts produced from the same gene has seldom been studied, for a number of reasons. First, not all of the protein products from the same gene are well understood. Second, a large number of alternative transcripts are expressed at very low levels or they are not stable; therefore sometimes they are ignored. Third, the regulation among the different transcripts from the same gene has not attracted much attention from the molecular biologists. To our knowledge, the present study may be the first to use mathematical modeling to study the regulation of different transcripts from the same gene.

These two transcripts have the same transcription initiation site at the 5’ end. They have the same 5’ untranslated region and the same first exon (counted from the 5’ end). However, the poly(A) sites at the 3’ end are different, with one located upstream of the other. The two poly(A) sites are called the upstream poly(A) site (UPA) and the downstream poly(A) site (DPA). If the UPA is used, then the DPA cannot be used. On the other hand, if the UPA is skipped, then the downstream poly(A) site may be used. The detailed mechanism of when and how the upstream or downstream poly(A) site is used is not known. The efficient usage of the poly(A) site is assumed to depend on the protein concentrations of the polyadenylation factors. Some polyadenylation factors have been found to affect the polyadenylation efficiency, such as CstF-64 [41]. Due to the limited understanding of the interaction between CstF-64 and CstF-77, we assume the concentration of CstF-64 is constant in our model. The CstF-77 long form protein is characterized as one subunit of CstF and is necessary for the cleavage of the pre-mRNA. The short form protein is assumed to either have no function or an inhibitory effect on the cleavage. If the short form has the same function as the long form, then it can be treated as a long form protein. We believe that the alternative poly(A) site may play an important role in regulating gene expression. By setting up the mathematical model and using numerical simulations, it is possible to deduce a putative function of the short form protein.

The regulation network can be described as a series of molecular events. When the transcription starts at the transcription initiation site, RNAP moves along the DNA sequence and the pre-mRNA is formed. When the UPA is transcribed, it can be used and the CstF-77.S is formed. If the UPA is skipped, then the RNAP continues to move and the DPA is then transcribed. When the DPA is used and the CstF-77.L is generated, the mRNA transcripts will be exported to the cytoplasm where they are translated into proteins. After folding of the protein and some modification processes, the proteins will then be transported back into the nucleus to act as factors to affect the transcription efficiency by polyadenylation.

DNA/RNA-protein interaction can be viewed as a substrate-enzyme interaction and the Michaelis–Menten equation has been used to describe this kind of kinetics [12]. Goodwin postulated an oscillatory model to simulate the dynamics of mRNA and protein [16, 17], and the dynamical behavior of this model was analyzed by Griffith [19, 20]. To set up the model, the following assumptions are made:

1. the higher the concentration of the long form protein present, the higher probability that polyadenylation occurs;
2. the short form protein has no impact or inhibitory effect on the polyadenylation;
3. the transcription initiation rate is constant, even though it is possible that no transcript may be generated for some transcription initiations.
Since the poly(A) site for the short form is located upstream of the poly(A) site for the long form, it is natural to assume that the transcription rates of the long and short forms, $P_l$ and $P_s$, respectively, have the following representations using the Goodwin oscillator [16]:

\[ P_l = I_0 - \alpha \frac{L^n}{K_s^n + \epsilon S^n + L^n}, \]
\[ P_s = \alpha \frac{L^n}{K_s^n + \epsilon S^n + L^n} \]

where $\alpha$ is the maximum transcription rate of the short form; $L$ and $S$ are the relative protein concentrations for the long and short forms, respectively; $I_0$ is the transcription initiation rate, which is naturally no less than $\alpha$; $K_s$ is the equilibrium constant; $n$ is the Hill coefficient; and $\epsilon$ is the relative inhibitory constant for the short form. Based on this setup, we can see that if $\epsilon = 0$, then the short form protein has no effect on the polyadenylation.

As we know, in eukaryotes, when a gene is turned on, a messenger RNA will be made, processed, and transported from the nucleus to the cytoplasm before the protein is generated. On the other hand, when a gene is turned off, the protein production will cease and the concentration of protein depends on the decay of the RNA that has already been transcribed. The delay time for the formation of a functional protein from a mature mRNA molecule has received some attention [9, 39]. The delay time may be different for different transcripts. The correlation between mRNA and protein abundance has been widely studied in other species [18, 34]; however, the correlation does not appear to be well defined because of noise and other factors. The mRNA can be measured by many techniques, such as microarray, polymerase chain reaction (PCR), etc. However, the protein concentration is relatively expensive to measure; therefore, the variables in the model will be chosen to be the mRNA expression levels. We assume that the protein level at the current time is proportional to the mRNA level that has been transcribed at some previous time. Thus the following relationships are assumed:

\[ L(t) \sim l(t - \tau_0), \quad S(t) \sim s(t - \tau_0) \]

where $t$ is the time variable; $l(t)$ and $s(t)$ are the relative mRNA expression levels of the long and short form at time $t$, respectively; $\tau_0$ is the time duration between the formation of the mRNA and the functional protein, including the time for mRNA transport from nucleus to cytoplasm, translation, protein folding, protein relocation, etc. Furthermore, we assume that the time delay is the same for both the long and short form proteins. So the dynamics of the long and short form transcripts can be modeled by the following differential-delay equations:

\[ \frac{dl}{dt} = I_0 - \alpha \frac{l^n(t - \tau_0)}{K_s^n + \epsilon S^n(t - \tau_0) + l^n(t - \tau_0)} - d_l l, \]
\[ \frac{ds}{dt} = \alpha \frac{l^n(t - \tau_0)}{K_s^n + \epsilon S^n(t - \tau_0) + l^n(t - \tau_0)} - d_s s \]

where $d_l$ and $d_s$ denote the degradation rates for the long and short form transcripts, respectively. For some transcription initiations, both UPA and DPA could be skipped,
and hence no transcript could be produced. The probability that the DPA is used is assumed to depend on the concentrations of the polyadenylation factors, characterized by the Hill equation with a different equilibrium constant as used for the UPA. By slightly modifying (2.4), we obtain a more complete and complicated system:

\[
\begin{align*}
(2.6) & \quad \frac{dl}{dt} = \left( I_0 - \alpha \frac{l^n(t-\tau_0)}{K^n_s + \epsilon s^n(t-\tau_0) + l^n(t-\tau_0)} \right) \frac{l^n(t-\tau_0)}{K^n_l + \epsilon s^n(t-\tau_0) + l^n(t-\tau_0)} - dl, \\
(2.7) & \quad \frac{ds}{dt} = \alpha \frac{l^n(t-\tau_0)}{K^n_s + \epsilon s^n(t-\tau_0) + l^n(t-\tau_0)} - ds,
\end{align*}
\]

where \( K_l \) is the equilibrium constant for the polyadenylation of the long form. The above two equations (2.6)–(2.7) can also be written as probabilistic equations as follows:

\[
\begin{align*}
(2.8) & \quad \frac{dl}{dt} = (I_0 - \alpha p_s) p_l - dl, \\
(2.9) & \quad \frac{ds}{dt} = \alpha p_s - ds.s
\end{align*}
\]

where

\[
(2.10) \quad p_l = \frac{l^n(t-\tau_0)}{K^n_l + \epsilon s^n(t-\tau_0) + l^n(t-\tau_0)}, \quad p_s = \frac{l^n(t-\tau_0)}{K^n_s + \epsilon s^n(t-\tau_0) + l^n(t-\tau_0)}.
\]

\( p_l \) and \( p_s \) are the probabilities that the long and short forms are produced given the same condition.

3. Theoretical analysis of DDEs. In general, differential-delay equations (DDEs) are difficult to analyze mathematically. A DDE, also called a retarded functional differential equation, is a special type of functional differential equation. It is very similar to an ordinary differential equation (ODE), but its solution involves past values of the state variables. Thus, the solution of a DDE requires knowledge of the current state, as well as the state of some variables at an earlier time. Therefore, the DDE is an infinite-dimensional system, and we need to specify an initial function defined in an initial time interval for the initial condition. To study a simpler two-dimensional system than (2.6)–(2.7), we set \( K_l = 0 \) and \( \epsilon = 0 \):

\[
\begin{align*}
(3.1) & \quad \frac{dl}{dt} = I_0 - \alpha \frac{l^n(t-\tau_0)}{K^n_s + l^n(t-\tau_0)} - dl, \\
(3.2) & \quad \frac{ds}{dt} = \alpha \frac{l^n(t-\tau_0)}{K^n_s + l^n(t-\tau_0)} - ds.s.
\end{align*}
\]

The above system assumes that the short form protein has no function and that every transcription initiation generates one transcript, either the long form or short form. Note that the first equation is independent of the second equation, so the above system can be studied by the single DDE (3.1). This equation has six parameters \( I_0, \alpha, n, K_s, d_l, \tau_0 \), and it can be reduced to five parameters by the following transformation:

\[
(3.3) \quad \hat{l}(t) = \frac{l(t)}{K_s}
\]

We then obtain

\[
(3.4) \quad \frac{d\hat{l}}{dt} = I_1 + \frac{\alpha_1}{1 + l^n(t-\tau_0)} - d_l \hat{l}
\]
where
\begin{equation}
I_1 = \frac{I_0 - \alpha}{K_s}, \quad \alpha_1 = \frac{\alpha}{K_s}.
\end{equation}

Dropping the \(^\hat{\cdot}\) on \(\dot{l}(t)\), we obtain
\begin{equation}
\frac{dl}{dt} = I_1 + \frac{\alpha_1}{1 + ln(t - \tau_0)} - dl.
\end{equation}

The above equation has been studied for \(I_1 = 0\) and some results were obtained in [28]. Here, \(I_1 \geq 0\) since \(I_0 \geq \alpha\). Equation (3.6) has only one steady state real solution \(\bar{l}_0\), which is independent of the time delay. The following questions can be asked: Does (3.6) have a periodic solution? If it does, what is the relationship between the period and the delay?

The stability of this steady state solution \(l(t) = \bar{l}_0\) can be found by linearizing the nonlinear term around the steady state. However, this calculation is much more difficult than for ODEs because of the time delay in the differential equation. After linearization, the following linear DDE is obtained:
\begin{equation}
\frac{d\hat{x}}{dt} = \hat{\alpha} \hat{x}(t) + \hat{\beta} \hat{x}(t - \tau_0)
\end{equation}
where
\begin{equation}
\hat{x}(t) = l(t) - \bar{l}_0, \quad \hat{\alpha} = -d_l, \quad \hat{\beta} = -\frac{\alpha_1 \bar{m}_0}{(1 + l_0^\mu)^2}.
\end{equation}

We further scale the time and obtain the following linear DDE by dropping the \(^\hat{\cdot}\) on \(\hat{x}(t)\):
\begin{equation}
\frac{dx}{dt} = \alpha x(\hat{t}) + \beta x(\hat{t} - 1)
\end{equation}
where
\begin{equation}
\alpha = \hat{\alpha} \tau_0, \quad \beta = \hat{\beta} \tau_0, \quad \hat{t} = \frac{1}{\tau_0} t.
\end{equation}
The above equation (3.9) was analyzed in [6, 10] and some of the main results will be briefly summarized here.

Set \(x(t) = e^{zt}\); then the characteristic equation of (3.9) is given by
\begin{equation}
z = \alpha + \beta e^{-z}.
\end{equation}
Writing \(z = \mu + i\nu\) where \(\mu\) and \(\nu\) are real and substituting into the characteristic equation, we obtain the parametric equations for \(\alpha\) and \(\beta\):
\begin{align}
\alpha &= \mu + \frac{\nu \cos \nu}{\sin \nu}, \\
\beta &= -\frac{\nu \sin \nu}{\sin \nu}.
\end{align}
Set \(\mu = 0\) to obtain the pure imaginary solutions of the characteristic equation, and we have
\begin{align}
\alpha &= \frac{\nu \cos \nu}{\sin \nu}, \\
\beta &= -\frac{\nu}{\sin \nu}.
\end{align}
Fig. 3.1. Parameter plane for the linear DDE. \( \alpha \) and \( \beta \) are the coefficients in (3.9). \( R^\pm \) are the regions bounded by the solid curves, and \( C^\pm \) are the solid curves generated by (3.14)–(3.15). The numbers given after the colon represent the number of roots of (3.11) located in the right-half plane.

The right-hand sides of (3.14)–(3.15) are even in \( \nu \), so we can restrict \( \nu \geq 0 \). There are singularities at \( \nu = k\pi \), \( k = 0, 1, 2, \ldots \). Therefore, we can separate the right-half axis of \( \nu \) into small intervals such that the sine function has a single sign. Define the following intervals:

\[
I_k^- = ((2k-1)\pi, 2k\pi), \quad I_k^+ = (2k\pi, (2k+1)\pi).
\]

The curves \( C_k^\pm \) in the \((\alpha, \beta)\)-plane are parameterized by \( \nu \) within each interval \( I_k^\pm \), as follows:

\[
C_k^\pm = \left\{ (\alpha, \beta) : \left( \frac{\nu \cos \nu}{\sin \nu}, -\frac{\nu}{\sin \nu} \right) \mid \nu \in I_k^\pm \right\}.
\]

For \( k = 0 \), \( I_0^- = (-\pi, 0), I_0^+ = (0, \pi) \), so \( C_0^- = C_0^+ = C_0 \). For all \( k > 0 \), \( C_k^+ \) and \( C_k^- \) are different curves. We can plot all the curves in the same \((\alpha, \beta)\)-plane as shown in Figure 3.1. The open regions \( R_k^\pm \) are the areas bounded by the solid curves. By analyzing the curves in detail, the following properties about the roots of the characteristic equation (3.11) can be observed [6, 10]:

1. there are two pure imaginary roots if the parameter pair \( \alpha \) and \( \beta \) is located on the curve \( C_k^\pm, k = 0, 1, \ldots \);
2. there is no (one) root in the right-half plane if the parameter pair \( \alpha \) and \( \beta \) is located in region \( R_0^+ (R_0^-) \);
3. there are \( 2k+1 \) (2k) roots in the right-half plane if the parameter pair \( \alpha \) and \( \beta \) is located on the curve \( C_k^- (C_k^+) \) and region \( R_{k-1}^- (R_k^+) \), \( k = 0, 1, \ldots \).
From (3.8) and (3.10), we know that $\alpha < 0$ and $\beta < 0$, which corresponds to quadrant III in the $\alpha \beta$-plane. The curves $C_k^+$ in quadrant III are generated by the parameter $\nu$ in the interval $(2k + 0.5)\pi, (2k + 1)\pi)$. If the system has a periodic solution, then it has the form $e^{i\nu t}$ where $\frac{\pi}{2} < \nu < \pi$. The period of the solution for (3.9), $T'$, is given by $T' = 2\pi/\nu$. Therefore, the period $T$ of the solution for (3.1) satisfies $2\tau_0 < T < 4\tau_0$. That is, if the time delay $\tau_0$ is specified, when the system has periodic solutions, the period should be between $2\tau_0$ and $4\tau_0$. Since we have very little information about the time delay, the relationship between the delay and the oscillatory period gives us a basic method for the estimation of the delay.

4. Numerical simulation. There are nine parameters in (2.6)–(2.7) and the solution could be very complex for some sets of parameters [6]. For simplicity, all of the parameters are assumed to be normalized. We have chosen some of the parameters according to the following considerations. For the transcription initiation rate, without loss of generality, we set $I_0 = 1$. The Hill coefficient representing the cooperative binding process is difficult to determine, and various values have been chosen by different researchers [19, 39, 50, 15, 11]. For this study, the Hill coefficient is chosen to be $n = 4$. Discrete delay times have been introduced to study gene regulation. Monk [30] used a transcriptional delay around 15–20 min to study the oscillatory expression of the transcription factors Hes1, p53, and NF-κB. Lewis, Warman, and Saunders [25] included a delay of 8 hr of the period protein to study circadian rhythms. From previous theoretical analyses, the period $T$ is estimated to be between two and four times the delay time if the model of the form (3.1) has a periodic solution. Also, this is almost true for system (2.6)–(2.7) if the system has a periodic solution resulting from some random simulations [6]. It has been found that the protein production rates fluctuate over a time scale of about one cell cycle [37]. For a typical mammalian cell, the cell cycle lasts about 24 hr. Thus, it is natural to estimate the time delay $\tau_0$ to be 8 hr, one third of the cell cycle period.

To solve the DDEs, we have to estimate the remaining six parameters. Generally, these parameters are difficult to estimate from biological experiments, and so an in silico estimation is made based on available experimental data. An optimization method will facilitate the estimation procedure. Delay differential equations are difficult to solve analytically [31], so we use the package dde23 in MATLAB [40] and DKLAG6 written in Fortran 77 [8] to solve the DDEs numerically. Here, we use a constant initial function (CIF) $l(t) = s(t) = 1$, where $-\tau_0 \leq t \leq 0$. To accelerate the estimation of the parameters, the message passing interface for parallel computing is adopted, and the computations are done using the Hydra Cluster in the Department of Mathematical Sciences at the New Jersey Institute of Technology, funded by a grant from NSF.

5. Experimental data. HeLa cells (an immortal cervical cancer cell line taken from Henrietta Lacks) were seeded at 70% confluence in a 12-well plate in a Dulbecco’s Modified Eagle Medium supplemented with 10% fetal bovine serum. Total cellular RNAs were extracted using the RNeasy kit (Qiagen) according to the manufacturer’s protocols. mRNAs were reverse-transcribed using oligo-dT primers (Promega). Real-time quantitative PCR was carried out using the 7500 Real Time PCR system (Applied Biosystems) with Syber-Green I as dye. The following primers were used for detecting various regions of the CstF-77 gene: F0 (5’-GAGGCCATGTCAGGAGAC), R1(5’-GCAACTCCAAAATGCA-ACAA), R3(5’-CATAAATCAATGTGCAAAACC). Primers for Cyclophilin A (CYPH) were 5’-ATGG-TCAACCCCACCGTGT and 5’-TTCCTGCTGCTTTTGAAACTTTGTC. Data were obtained at the five time points
Fig. 5.1. Hypothetical experimental data at five time points. The x-axis is the time in hours and the y-axis is the relative expression level normalized to the first time point at 16 hr. The solid (dashed) line is the relative expression of long (short) form. The error bar is based on two experimental repetitions.

16, 24, 36, 48, 56 hr, and the relative expression levels are shown in Figure 5.1. However, for the theoretical work, we use these as hypothetical data for our model since the two experiments measuring the long and short form were done at different times and presumably they are not comparable. Therefore, we assume that there exists a periodic oscillation in the experimental data. The oscillation has also been confirmed in our most recent unpublished manuscript submitted for recent publication [27]. The exact periodic oscillation is difficult to capture due to experimental limitations.

6. Parameter estimation. The parameters in the model equations are estimated by minimizing the error between the experimental data and the model simulation data. Assume the experimental data contain \( m \) time points at \([t_1, \ldots, t_m]\) for both long and short form mRNA expression levels, i.e., \( E = [l_1, \ldots, l_m; s_1, \ldots, s_m] \). For some given parameter values, the DDE model system can be solved and evaluated at these \( m \) time points, i.e., \( N = [\hat{l}_1, \ldots, \hat{l}_m; \hat{s}_1, \ldots, \hat{s}_m] \). The mean square error is calculated by

\[
err = \sqrt{\frac{\sum_{i=1}^{m} [(l_i - \hat{l}_i)^2 + (s_i - \hat{s}_i)^2]}{m}}.
\]

We want to find an optimal parameter set to minimize the above error. Very little information is known about the magnitude of these parameters, so we assume that all of the parameters are located between 0 and 1. In fact, with \( I_0 = 1 \) and constant initial conditions, these assumptions are reasonable based on the biological meanings of the parameters. To find an optimal parameter set in the six-dimensional parameter space, we use a global searching method. We evenly divide each parameter axis into \( p \) points and by estimating all of the possible combinations of the parameter values, we can find the global minimal error by solving the DDEs \( p^6 \) times. Using the DKLAG6
package, it takes about 0.0011 sec to solve the DDE once, so the entire parameter estimation calculation will take about 0.0011 p6 sec. For example, if \( p = 50 \), i.e., we search the space with step 0.02, it would take half a year to finish the calculation. To accelerate the computation, we used 20 nodes from the NJIT Hydra Cluster and it would take about one week to finish. To further accelerate the computation, we assume the degradation rates for the long and short form are the same, i.e., \( d_l = d_s \). In this case, it took 2 hours to complete the job with \( p = 50 \), 20 cluster nodes, and five unknown parameters, which we found to be acceptable. Another reason for using the global search method, instead of an heuristic optimization method, such as a genetic algorithm and simulated annealing, is that we also want to estimate the variance and the distribution of the parameters.

7. Results. The optimal parameter values from the optimization are

\[
\alpha = 0.88, K_s = 1.00, d_l = d_s = 0.44, \epsilon = 0.02, K_l = 0.02,
\]

which led to the smallest value of \( err = 0.43 \) with CIF of \( l(t) = s(t) = 1 \) for \( -\tau_0 \leq t \leq 0 \). If not stated otherwise, we always used the above CIF. The simulation results and the experimental results are plotted in Figure 7.1, from which we can see that the simulated oscillation gives a good qualitative fit to the experimental data. In addition to the optimal parameter set, we also saved all of the suboptimal parameter sets with the mean square error less than 0.5. These additional suboptimal parameters are plotted in Figure 7.2 to show their distributions.

![Fig. 7.1. Solutions fitted with the optimal parameter values. The optimal parameter set is obtained by the minimization of the error between the hypothetical experimental data and the simulation results. The oscillatory period is estimated to be around 22 hr. The experimental data are shifted 100 hr to the right for better representation.](image)
Several observations can be made from the optimized parameter values.

1. Parameters $\alpha, K_s, d_l(d_s), \epsilon$ have smaller variance than $K_I$. This may tell us that $K_I$ is less critical in determining the model behavior.

2. Parameter $\alpha$ is close to 1. This indicates that if the long form protein is overexpressed, then the cell would generate more short form transcripts. The production rate of the short form transcript approaches the transcription initiation rate to reduce the overall redundancy of the long form protein.

3. $K_I$ is less than $K_s$. This implies that if the long and short form transcripts both have the possibility for polyadenylation, then the probability of generating the long form transcript is higher than the short form. The reason for this might be the fact that the long form transcript is functional and the short form is nonfunctional.

4. The parameter $\epsilon$ is very small. Since we assumed that the short form could have either no function or an inhibitory function, then we can hypothesize that the short form protein has no function. The CstF-77.S protein has not yet been detected by experimental biologists. If expressed, then from the putative protein sequence, it would only contain the first HAT domain, and lack the C-terminal region responsible for interacting with CstF-64, CstF-50, and itself [44], which confirms that $\epsilon$ should be close to zero.

Knockdown or knockout biological experiments are usually taken to test some biological models and hypotheses. Here, we can also do some numerical simulations for these “knockdown” or “knockout” experiments by changing the initial condition of the differential equations. For example, if the long form is knocked down, the CIF for that long form could be set to a smaller value. We tried different CIFs and simulated the essential “phenotype.” From the simulation, we observed the following.

1. If the long form transcript is knocked out, i.e., the corresponding initial condition of $l(t)$ is set to zero or to a very small value, such as less than 0.06 for our optimized parameter, then the solution goes to zero; see, Figure 7.3(a). Biologically, this implies that the cell will die without a minimum amount of long form protein. It has been confirmed by experiment that if the suppressor of the forked gene ($Su(f)$), a Drosophila homologue of $CSTF3$, is knocked out, the Drosophila will die at the larvae stage [42].
2. If we only knock down the long form, with the initial condition of \( l(t) \) set to a small value but greater than some threshold, e.g., 0.07, the cell can recover very slowly and then survive; see, Figure 7.3(b). This means that knocking down the long form would not kill the cell, and a small amount of the long form protein is sufficient for the cell to rescue itself.

3. If the long form is overexpressed, i.e., setting the CIF for \( l(t) \) to 5, the system returns to a final oscillation; see Figure 7.3(c). Overexpressing or knocking down the short form does not affect the system too much except for the initial perturbation; see Figure 7.3(d)–(e).

4. From Figure 7.3, we can see that during some period, the number of long and short form transcripts both increase or decrease. This agrees with the results that the lack of the Su(f) function is correlated with the disappearance of the short form Su(f) RNA and accumulation of the short form requires the wild-type Su(f) protein [1].

The set formed by all the initial points in the phase space of a dynamical system, which is attracted to a given solution (e.g., a fixed point or a limit cycle), is called the basin of attraction of that solution. The boundaries of the various basins of attraction are known as basin boundaries. The dependence of the solution behavior on the initial conditions in certain first-order nonlinear DDEs has been investigated [22]. From Figure 7.3(b)–(e), the system may converge to the same stable periodic solution for different initial conditions. We hypothesize that this is a limit cycle solution. It is easy to see that the origin is a fixed point of the system (2.6)–(2.7). By linearizing the system around the origin, we can find that the origin is a stable fixed point. However, we don’t know if the limit cycle solution is also stable.
Fig. 7.4. Bifurcation solutions of (2.6)–(2.7) with different CIFs. The long form CIF bifurcation point with accuracy to 11 digits is generated so that the solutions go to the origin with the given CIFs and go to the limit cycle solution when the last digit increases by 1. (a) $l_0 = 0.0041328205677$, $s_0 = 0.01$. (b) $l_0 = 0.065228128392$, $s_0 = 1.0$. (c) $l_0 = 6.4525236006$, $s_0 = 100$. (d) Small oscillatory solutions of three different CIFs. The solutions of the long form would be the same except for a small shift.

We notice that there is a bifurcation point of the long form CIF if we fix the short form CIF at some constant; see Figure 7.4. For example, if the initial condition for the long form is greater than 0.0653 with the short form fixed at 1, the system returns to the limit cycle. Otherwise, it will go to zero. For different short form CIFs, the bifurcation points for the long form can be estimated numerically and they are plotted in Figure 7.5. The curve in the middle is the basin boundary, separating the phase space into two regions: zero region and oscillatory region. If the initial condition starts in the zero region (oscillatory region), the solutions are attracted to the origin (limit cycle solution). We found two interesting phenomena about the basin boundary: when the initial conditions are far away from the origin, the basin boundary is almost a straight line with the ratio (long/short) approaching 0.064; when the initial conditions are close to the origin, the basin boundary becomes a vertical line. That is, if the long form is greater than some small value such as 0.00413, the solution always goes to the limit cycle, independent of how small the short form is. This could be explained biologically: the short form protein could be devoid, but the small amount of long form protein is necessary to rescue the cell.

It is not clear why the basin boundary approaches the line with ratio (long/short) around 0.064 if the initial conditions are far away from the origin. For three different initial conditions of the short form, $s_0 = 0.01, 1, 100$, we calculated the corresponding long form bifurcation initial conditions with the precision to 11 digits, $l_0 = 0.0041328205677, 0.065228128392, 6.4525236006$, as shown in Figure 7.4(a)–(c). With the initial condition $(s_0, l_0)$, the solutions go to zero. When the last digit of $l_0$ increases by one, the solutions go to the limit cycle solutions. From Figure 7.4(d),
Fig. 7.5. Dynamical behavior of (2.6)–(2.7) with optimal parameters. The zero region is to the left of the basin boundary, and the oscillatory region is to the right of the basin boundary. Three fixed points are indicated by the small blue circles. If the CIF starts in the zero region, the solution is attracted to the origin; if the CIF starts in the oscillatory region, the solution is attracted to the limit cycle solution; if the CIF starts exactly on the basin boundary, the solution for the long form transcript oscillates with a small amplitude, and the solution for the short form transcript goes to zero rapidly.

We hypothesize that for given short form CIF $s^*$, we can always find $l^*$ such that the solution of (2.6)–(2.7) with the initial condition $(l^*, s^*)$ would be the long form oscillating with very small amplitude and the short form approaching zero. We call this solution the small oscillatory solution. With different $(l^*, s^*)$, the solutions of the long form would be the same except for a small shift (Figure 7.4(d)). We also noticed that the small oscillatory solution has the period $T \approx 9$, a slightly larger value than the delay time $\tau_0 = 8$, but much smaller than the period of the limit cycle solution. The justification of this numerical finding can be seen as follows.

When the small oscillation occurs in Figure 7.4(d), $s(t)$ goes to zero and $l(t)$ is very small with maximum magnitude around 0.005. $K_s$ is very large compared with $l(t)$, so (2.6) can be approximated by the following reduced equation:

\[
\frac{dl}{dt} = I_0 \frac{l^n(t - \tau_0)}{K^n_l + l^n(t - \tau_0)} - dl.
\]

When we linearize the above equation around the stationary point, we obtain the following linear DDE:

\[
\frac{d\hat{x}}{dt} = \hat{\alpha}x(\hat{t}) + \hat{\beta}x(\hat{t} - \tau_0)
\]

where $\hat{\alpha} < 0$ and $\hat{\beta} > 0$. Therefore, $(\hat{\alpha}, \hat{\beta})$ is located in the second quadrant of the parameter plane and correspond to the curve parameters in the interval $((2k + 1.5)\pi, (2k + 2)\pi)$. So the period is between $\tau_0$ and $\frac{4}{3}\tau_0$ and it agrees with the numerical simulation.
Fig. 7.6. The coefficients of the linear DDE (3.9) of (2.6) at point $P_3$. The circle represents the coefficients $(\alpha, \beta)$ of $P_3$. The curve above the circle corresponds to the curve $C^+_0$ in Figure 3.1 and below is curve $C^+_1$.

Furthermore, set $s(t) \to 0$, then the right-hand side of (2.6) becomes

\begin{equation}
(7.4) \quad f(l) = \left( I_0 - \alpha \frac{ln}{K^n + l^n} \right) \frac{ln}{K^n + l^n} - dl.
\end{equation}

Setting $f(l) = 0$ and solving for $l$, we get $\bar{l}_0 = 0.0041316256293$ and have $f'(\bar{l}_0) > 0$. This agrees very well with the numerical finding where the bifurcation point of the long form is close to 0.00413 when the short form is fixed at a very small value.

In addition to the two fixed points close to the origin as shown in Figure 7.5, there is one more fixed point at $P_3(1.0906, 1.1547)$, which is away from the origin and located in the oscillatory region. With $P_3$ as the initial condition, we should get the constant steady state solution. However, due to numerical noise, the solution goes to the limit cycle solution. This implies that $P_3$ is not stable. A simple proof can be given as follows. We linearize (2.6) around the point $P_3$ with the short form fixed and plot the normalized parameters in the $(\alpha, \beta)$-plane; see Figure 7.6. We notice that the coefficients $(\alpha, \beta)$ of the linear DDE (3.9) are located in region $R^+_1$ in Figure 3.1, and therefore, the characteristic equation (3.11) for $z$ has two roots in the right-half plane. Thus, point $P_3$ is unstable.
8. Conclusions and discussion. In summary, a mathematical model has been proposed to simulate the regulation between two different transcripts from the gene CSTF3. The model is able to produce solutions that approximate the hypothetical experimental data. The oscillatory solutions from the model can also explain why sometimes the long form mRNA cannot be detected even if the protein is necessary for the cell. This model can also be applied to other genes with similar polyadenylation patterns such as poly(A) polymerase α (PAPα). Multiple PAPα isoforms (I, II, III) have been found in human and mouse, and autoregulation may also exist in their expression [24, 51].

As the initial conditions change over a large area, the system shows a robust and stable oscillation. Robust means that the precise initial conditions are sometimes not critical for the oscillation to occur. It also implies that the polyadenylation process is very stable, suggesting its importance in the cell cycle. The dynamical system with the optimized parameters has been studied in detail. The initial conditions of the model lead to the bifurcation of the stable solutions, which agrees well with the fact that knocking out the polyadenylation factors would kill the cell and some amount of long form is essential to rescue the cell. The basin boundary has been found numerically, and it shows a very interesting pattern. Three basins of attraction have been found: limit cycle solution, zero solution, and small oscillatory solution with the short form approaching zero and long form approaching a very small amplitude oscillation.

By adding more factors, such as CstF-64, the model can be extended to more complex biosynthetic pathways. It still remains to identify the complete alternative polyadenylation mechanism biologically. This mathematical model provides molecular biologists with a new viewpoint of the hypothesized mechanism and indicates the importance of the nonfunctional transcript due to alternative polyadenylation. A nonfunctional transcript can also play an important role in gene regulation by generating nonfunctional products. This study provides a comprehensive and consistent mathematical framework for understanding the alternative polyadenylation process.

Acknowledgments. We thank Amitabha Bose from the New Jersey Institute of Technology for valuable discussions about the dynamical system and Sylvester Thompson from Radford University for help on the usage of the DDE solver.

REFERENCES


