Proteomic waves in networks of transcriptional regulators

A. S. Carstea

Horia Hulubei National Institute for Physics and Nuclear Engineering (IFIN-HH), Dept. of Theoretical Physics 407 Atomistilor, Magurele - Bucharest, 077125, Romania

The protein production for a gene regulatory network model with activation-repression links is analysed. In these networks this production depends on how proteins induce or repress the genes. Experiments show that networks of inducers or repressors exhibit bistability or oscillatory behaviour of protein production. Here we report a completely novel aspect, namely for different promoter activity functions, protein production (initially localised on a certain number of genes) can propagate to the others in a “solitonic” way. In particular, the chemical rate equation for the whole network can be solved exactly and in the case of big number of operator sites the proteomic signal along the gene network is given by a superposition of perturbed dark solitons of defocusing semidiscrete modified Korteweg de Vries equation.

PACS numbers: 05.45.Yv, 87.14.Gg, 82.39.Fk

Genetically precoded responses of all living organisms to external perturbations and signals are achieved through networks of genes of high complexity. The genes represent coded segments of the DNA-chain, any code being a sequence of base pairs $S = \{b_1, b_2, ..., b_n\}$ where any $b_i$ belongs to the set $\{A, G, C, T\}$. The role of the genes (crucial for any living organism) is protein production as an outcome of genetic transcription-translation process. This one consists of three steps: In the first, the RNA polymerase (RNAP) sticks to the gene promoter (the starting position of the gene) and, with the aid of some proteins called transcription factors (TF), the gene is “read” by RNAP as it is moving along the DNA. In the second, during this “travel” RNAP produces a big molecular chain the messenger RNA (mRNA) copying the code. Finally, the mRNA goes to the ribosomal machine which produces the protein according to the codon distribution of the gene. The level of gene expression is basically the quantity of protein produced by gene itself. Proteins are responsible for all processes in living organisms (metabolism, membrane channels etc.) but, as we said above there are some which can activate or repress the transcription process, namely transcription factors. Typically a gene is subject to the regulatory effect of other genes which can act on it (by means of TF) in either an activating or a suppressing way, depending on the situation [1]. Sometimes it is useful to describe it in a digital way, namely a gene is in the state “ON” if it produces the protein (i.e. it is active) and in the state “OFF” if it is supressed (or its own protein production is low) [2].

A good understanding of the dynamics will allow to design specific artificial gene regulatory networks performing certain functions. The problem of design is quite complicated due to the poor understanding of the “design principles”. Up to now a large amount of studies have been performed in vitro and in vivo to understand the molecular mechanisms of some simple genetic circuits [3]. More precisely, experiments with networks of few activators or repressors display many stable sates or oscillatory behaviour with respect to protein production[4]. The theoretical approach relies mainly on the study of rate equations for concentrations of various chemicals involved in the network dynamics (mRNA, TF, etc.).

In this paper we are going to show using the differential rate equations that under very simple assumptions a big chain of transcriptional regulators having inducer-repressor design can display complex and in the same time controlable dynamics. Namely, the gene expression can “propagate” to the other genes in a “solitary wave” manner and as a readout the concentrations of transcription factors can have ”solitonic” dependence with respect to the genes. More precisely what are we going to show is the following result: Suppose that at a certain initial moment there is a distribution of (TF) for every gene in the network. This distribution is to be taken as the initial condition for an “evolution equation” [5]. This evolution equation is nothing but the whole system of differential rate equations for every gene. For big number of genes this system can be considered as a partial differential-discrete nonlinear equation (which in our model turns out to be of Korteweg de Vries-type). Accordingly the solution will give the evolution in time of (TF)-distribution for every gene.

Chemical rate equations describe the transcription-translation mechanism. In order to build a mathematical model of this process, one has to describe first the binding of the RNAP molecule to the DNA promoter region (which is at the beginning of the gene). In a thermodynamical description [6], the promoter activity is proportional to the equilibrium probability $g_{A,R}(TF)$ of the binding of the RNAP to the core promoter. In the case of the simple activation or suppression, the dependence on the cellular TF concentrations (which we shall denote by $p$) is described by the Arrhenius form [7]

$$g_A(p) = \frac{1 + \omega_A(p/K_A)}{1 + p/K_A}$$
The perspective operator sequence in the regulatory region and is the effect of promoter leakage in repression. In a \( K \) "stier" sigmoid (Fig.1),

\[
\text{Boltzmann weight of the activator-RNAp interaction,}
\]

the binding probability has a more general form as a Boltzmann weight

\[
\frac{1}{1 + p/K_R} + \Lambda
\]

for activation and

\[
g_R(p) = \frac{1}{1 + p/K_R} + \Lambda
\]

for suppression where \( \omega_A = \exp(-\Delta G_A - p/RT) \) is the Boltzmann weight of the activator-RNAp interaction, \( K_A \) dissociation constant between the protein and respective operator sequence in the regulatory region and \( \Lambda \) is the effect of promoter leakage in repression. In a more general context a detailed analysis of the interaction of RNA polymerase with the DNA chain shows that the binding probability has a more general form as a "stiffer" sigmoid (Fig.1),

\[
g_{\alpha,\beta}(p) = \frac{\alpha + \beta p^\sigma}{1 + p^\sigma}
\]

where \( \alpha \) and \( \beta \) are strictly positive numbers (for instance \( \beta/\alpha \) in the activator case is proportional to the Boltzmann weight \( \omega_A \) and \( p \) is normalised by dividing it to the dissociation constant \( p \rightarrow p/K_A \)). Activation occurs for \( \alpha < \beta \) whilst repression for \( \alpha > \beta \). Experimentally in the activator case \( 10 < \beta/\alpha < 100 \) and in the repressor case is slightly bigger [1]. The exponent \( \sigma \) is the Hill coefficient which can take only nonnegative values. Biologically it represents the cooperativity in the promoter activity and is related to the number of operator-bound transcription factors interacting with RNAp. It will parametrize our promoter activity functions.

Next step is the protein production from the mRNA which enters in the ribosomal machine. The ribosomal activity function will have a linear form,

\[
f(m) = \nu m - \mu
\]

where \( \nu \) is the protein synthesis rate at full activation and has a large span of values (0-100nM/min). The other term \( \mu \) is related to the fact that a certain amount of mRNA entering in the translation process is not coded [8].

From these basic ingredients, we can write the chemical rate equations for one gene. In the first one the RNA polymerase produces RNA-messenger acid \( m \)

\[
dm/dt = g_{\alpha,\beta}(p) - \lambda m
\]

where \( \lambda m^{-1} \) is the mRNA half-life which is around 5 min. Next, the RNA-messenger acid goes to a ribosomal machine and TF proteins are produced according to the equation:

\[
dp/dt = f(m) - \gamma p = \nu m - \mu - \gamma p
\]

where \( \gamma^{-1} \) is the protein half-life which is ten times higher than mRNA one.

Since the kinetics of RNA messenger production is rapid compared to those of the TF proteins, it is not unreasonable to make a steady-state assumption in the first equation, and thus we have \( m = g(p)/\lambda m \). Absorbing the ratio \( \nu/\lambda m \) into a redefinition of \( \alpha \) and \( \beta \) one can write the equation for the transcription factor production associated to a single gene:

\[
dp/dt = g_{\alpha,\beta}(p) - \mu - \gamma p
\]

FIG. 1: Promoter activity as function of TF concentration. The light blue curve represents the activator case and the dark blue one the repressor case. The Hill coefficient is 4 and the ratio \( \beta/\alpha \) is 10:1 on the g-axis

Now we can formulate our model. Consider a gene network where each gene is in interaction with two others, the effect of which can be either activating or suppressing. The genes are indexed with integers \( n \). The promoter activity functions are considered to have the same parameters \( \alpha \) and \( \beta \). The underlying network mechanism is the following: the protein \( p_{n-1} \) produced by the gene \( (n - 1) \) represses the gene \( n \) and the protein \( p_{n+1} \) produced by the gene \( (n + 1) \) activates the gene \( n \). This scenario works for every gene (i.e. every \( n \)). In addition we consider that the proteins \( p_{n-1} \) and \( p_{n+1} \) act independently on the regulatory region of the gene \( n \) and do not interact each other (which is not always true [6],[9]).

From the one-gene model presented above, we can write the equation (2) for any gene \( n \) as:

\[
\frac{dp}{dt} = g_{\alpha,\beta}(p_{n+1}) + g_R(p_{n-1}) - \gamma(n)p_n
\]

where \( g_{\alpha,\beta}(p_n) \) is given by the equation (1).

We consider this equation as an evolution partial differential-discrete one for the protein production \( p(n,t) \) of the whole chain. Our discuss will be focused on how an initial protein distribution localised on some genes evolves in time. This evolving distribution will be called proteomic signal or proteomic wave.

Assuming \( \alpha < \beta \), network equation will be:

\[
\frac{dp}{dt} = \frac{\alpha + \beta p_{n+1}^\sigma}{1 + p_{n+1}^\sigma} + \frac{\beta + \alpha p_{n-1}^\sigma}{1 + p_{n-1}^\sigma} - \mu - \gamma(n)p_n
\]
Here we have a strongly nonlinear differential-difference equation which gives the distribution in time of the transcription factors for all the genes. In order to put it into a more tractable form we are going to use the following substitution (valid since $p_n(t)$ is always positive)

$$p_n = e^{2\phi_n/\sigma}$$

And consider that $\beta = \alpha + L$ where $L$ is a positive number. Then after scaling time with $L$ the equation will take the form:

$$\frac{d}{dt} \left( e^{2\phi_n/\sigma} \right) = (\tanh \phi_{n+1} - \tanh \phi_{n-1}) + (2\alpha - \mu + L)/L - (\gamma(n)/L)e^{2\phi_n/\sigma}$$

Since $\alpha$ and $\gamma(n)$ are small and $L$, $\mu$ are big, the above equation can be seen as being:

$$\left( e^{2\phi_n/\sigma} \right)_t = \tanh \phi_{n+1} - \tanh \phi_{n-1} + \text{perturbation}$$

In order to see the structure of solutions we are going to analyse the above equation without perturbation for different $\sigma$'s. For $\sigma = 1$ putting

$$\phi_n = \frac{1}{2} \log \left( \frac{1}{w_n} - 1 \right), \quad 0 < w_n(t) < 1$$

one gets the celebrated integrable differential-difference Korteweg de Vries (KdV) equation

$$\dot{u}_n = u_n^2 (u_{n+1} - u_{n-1})$$

It is known that it admits multisoliton solution for arbitrary number of solitons and any value of wave parameters. Solitons are strongly localised travelling structures quite stable at interaction having amplitude-dependent velocity. They can be bright-type (bell-shaped) or dark-type (hole or well-shaped having finite asymptotic values). If TF-distribution is initially strongly localised on some genes then it will be described by superpositions of some genes then it will be described by superpositions of all the genes. In order to put it into some genes then it will be described by superpositions.

Accordingly, for $\sigma = 1$ possible solutions for proteomic signal are in the nonsolitonic sector.

In the case $\sigma > 1$ the phenomenology we expect to be richer [5]. Making an expansion in the left exponential and rescaling time we obtain:

$$\frac{d\phi_n}{dt} = \tanh \phi_{n+1} - \tanh \phi_{n-1}$$

and with the substitution $\phi_n = \tanh^{-1} w_n$ with $-1 < w_n < 1$

$$\dot{w}_n = (1 - w_n^2) (w_{n+1} - w_{n-1})$$

which is the defocusing differential-discrete modified Korteweg de Vries (mKdV) equation. This equation which has been analysed in detail in [11] admits only dark-type soliton as localised solutions.

Assuming that the positive nonzero boundary is $0 < w_0 < 1$ the 1 dark soliton has the following rather complicated form:

$$w_n = \frac{G_n F_n - G_{n+1} F_{n+1} - G_{n-1} F_{n-1}}{G_{n+1} F_{n+1} - G_n F_n}$$

where

$$G_n(t) = 1 + ae^{kn - \omega t}, F_n(t) = 1 + e^{kn - \omega t},$$

and $\omega = 2(1 - w_0^2) \sinh k$,

$$a = \frac{(1 + w_0) \exp(-k) + (1 - w_0) \exp(k) - 2}{(1 + w_0) \exp(k) + (1 - w_0) \exp(-k) - 2}$$

Accordingly, the proteomic signal given by,

$$p_n(t) = 1 + \frac{2}{\sigma} \tanh^{-1} w_n$$
FIG. 2: The protein distribution for the range of genes from 4 to 20. The genes 9, 10, 11 are in the repressed (OFF) state, all others being in active state and this pattern is propagating in time and backward with respect to the gene axis. Mathematically it corresponds to 1-dark soliton solution.

It is positive and the singularity region is avoided provided $|k| < \log((1 + w_0)/(1 - w_0))$. Moreover, the integrable character of the equation exhibits general dark multisoliton solutions which can be constructed as Hirota bilinear formalism usually does [10]. These dark-solitons can be seen as a propagation of proteomic signal for some genes in lower "gene expression" state (OFF-state) in a background of genes in higher state (Fig.2). Of course the ON/OFF picture is approximative since the level of gene expression is given by the soliton amplitude (depth in the dark case) in different points. The interesting fact is that the bigger is the amplitude (i.e. the closer is $k$ to $\log((1 + w_0)/(1 - w_0))$) the wider is the soliton shape. Moreover, when two different amplitude solitons collide the deeper one swallows the other and during the interaction the last one is turned into a bright one (reversed polarity). It means in the OFF region appear some genes in the state ON (Fig.3). This picture occurs for an arbitrary number of dark solitons in interaction. In addition $k$ is completely free (in its region) so is the amplitude (gene expression), width and velocity. The presence of perturbation will vary the amplitude and wave numbers in time. Another interesting aspect is that the equation does not admit bright solitons as localised solutions. It means that an initial distribution of type ON (all other genes being in state OFF) cannot propagate in a solitary way. Thus the only "solitonic" phenomenology appears to hold only for dark case.

Biologically, these solitons are propagating in "space of genes" since $n$ is the index of the gene and it can be in any position subject only to the interaction with specific transcription factors generated by the genes $(n + 1)$ and $(n - 1)$.

In conclusion, the distribution of TF concentrations is given by the solution of a perturbed nonlinear semidiscrete evolution equation which, for $\sigma = 1$ (one operator site) is KdV in the nonsolitonic sector and for multi-operator sites ($\sigma >> 1$) is the defocusing mKdV in the dark-soliton sector. Many interesting things arise from this model. First of all all the parameters involved in our equations are tunable($\alpha, \beta, \nu$) [6]. As an outcome one can in principle control de genomic signal propagation. Moreover the integrable character of the equation shows that the results are valid for a large domains of wave numbers ($k$) in the solitonic solutions. So, in principle one can start with many possible strongly localised initial conditions (given by genes in state OFF) and to expect a nice dynamics. Of course from the experimental point of view this is extremely difficult but this is exactly what one wants from a model - to exhibit a rather stable dynamics and to have trains of proteomic signals from various localised initial conditions. Even though the experimentators work with a few number of genes we expect that in the future will be possible to construct big modules with big number of genes in interaction. Our model will serve as a possible robust transmission network one similar to the solitonic fiber optical devices. Of course many open questions remain. First of all we did not consider the stochastic effects in the production of mRNA [12]. But working with big number of genes and implicitly with high concentrations the stochastic effects can be negligible. Also we have neglected the time-delay in the gene equations and possible polymerisations of transcription factors before entering into regulatory regions. We intend to consider these problems in a future publication.


