Title

Computational mechanisms in genetic regulation by RNA

Author

Deutsch, JM

Publication Date

2018-12-01

DOI

10.1016/j.jtbi.2018.09.016

Peer reviewed

Computational mechanisms in genetic regulation by RNA

J. M. Deutsch

Department of Physics, University of California, Santa Cruz CA 95064

Abstract

The evolution of the genome has led to very sophisticated and complex regulation. Because of the abundance of non-coding RNA (ncRNA) in the cell, different species will promiscuously associate with each other, suggesting collective dynamics similar to artificial neural networks. A simple mechanism is proposed allowing ncRNA to perform computations equivalent to neural network algorithms such as Boltzmann machines and the Hopfield model. The quantities analogous to the neural couplings are the equilibrium constants between different RNA species. The relatively rapid equilibration of RNA binding and unbinding is regulated by a slower process that degrades and creates new RNA. The model requires that the creation rate for each species be an increasing function of the ratio of total to unbound RNA. Similar mechanisms have already been found to exist experimentally for ncRNA regulation. With the overall concentration of RNA regulated, equilibrium constants can be chosen to store many different patterns, or many different input-output relations. The network is also quite insensitive to random mutations in equilibrium constants. Therefore one expects that this kind of mechanism will have a much higher mutation rate than ones typically regarded as being under evolutionary constraint.

1. Introduction

The overwhelming majority of transcripts in the human genome produce non-coding RNA (ncRNA) and these have been under intensive investigation in recent years [1, 2, 3, 4, 5, 6] which has revealed many functions. However, research to date has still only scratched the surface of the mechanisms involving these transcripts.

Aside from specific mechanisms, it is useful to take a step back and ask at an algorithmic level, what all of this extra RNA might be capable of doing, given the constraint that the mechanisms be biologically plausible. The author proposed [7] that a general way of understanding many of ncRNA's functions was to have these molecules act *collectively*. By collective behavior, what is

Email address: josh@ucsc.edu (J. M. Deutsch)

meant is that the actions of any one piece of the genomic circuitry is influenced by a large number of different molecules. This contrasts with the usual way of understanding biological regulation, where specific molecules will interfere, suppress, or promote, gene expression. This is most often how elements in *cis*-regulation are described. With collective mechanisms, such specific pathways cannot explain function. The system needs to be considered in its entirety for the correct genomic behavior to emerge.

The "connectionist" model of human cognition and machine learning, has been considered in rather early work in the context of many biological processes including gene regulation [8]. One can model regulatory elements physically, where binding and unbinding are controlled by equilibrium constants. The binding of regulatory proteins in such networks is generally thought to be quite specific, and these models are more akin to circuit diagrams with a few connections in and out of every element. The program pursued here is to understand if it is biologically tenable to instead have a large number of molecules present that have much less specific interactions, and yet can function in a precise way, regulating many of the myriad functions that take place in the cell. Of course, this is not meant to suggest that all functions operate this way, but that this collective mechanism could also be operating.

Recent work on "genome-wide association studies" (GWAS) [9, 10], have shown that a large number of traits are determined by the combined effect of many single nucleotide polymorphisms (SNPs). For example, the heritability of height involves more than 3×10^5 SNPs [11]. Typically in these cases, each SNP has a minute effect and it is the collective action of all of them that is largely responsible for the observed correlation of the trait with the genome. Therefore, at the least, the idea of such collective regulation is not inconsistent with what is known about genetics.

This paper describes a surprisingly simple mechanism for achieving this kind of collective regulation, where perhaps thousands of RNA species bind to each other promiscuously, yet this results in, or indeed is responsible for, a high level of computational complexity. This is motivated by the developments in artificial intelligence that have come about from consideration of similar collective models [12]. One of the most general kinds of models in this class is the "Boltzmann machine" [13], which in a certain limit, described below, becomes the so-called "Hopfield model" [14, 15, 16, 17, 18, 19].

Earlier work by the author [20] described a way of mapping the Hopfield model onto promiscuously binding RNA molecules. However it was not clear how such a mapping could be made compatible with the biological and physical requirements. The work here broadens the class of physical systems, and also shows how the mechanism can be greatly simplified to make it much more credible that it would have been able to evolve. Other recent work [21] proposes more direct methods for constructing chemical analogies of Boltzmann machines, but so far it is not clear how this is related to biology.

To make the proposal here biologically plausible, its mechanisms should involve functions similar to those already known to exist. There are two mechanisms necessary in what follows in order for it to work. The first is that different

chemical species bind and unbind in accord with statistical mechanics. The second is the existence of molecular mechanisms to selectively transcribe species of RNA depending on the fraction of it that is bound to other species. There are many ways of achieving the right mathematical form and there are many forms for this dependence that will work. This sort of behavior is fairly typical in many biochemical subsystems. A speculative proposal for accomplishing this would be that this process takes place with little genomic involvement. There is machinery capable of replicating RNA, similar to RNA-dependent RNA polymerase (RdRP), which is essential for the viability of many viruses [22], but also appears to exist in humans [23]. Furthermore the transcription rate of RdRP should depend on the relative amount of bound to unbound polymer for every molecular species. A more conventional approach uses the effect that ncRNA has on genomic transcription factors. This general kind of mechanism has been observed [24] in different situations. These possibilities are discussed in Sec. 6.2.

One of the main points of this work is to illustrate that there may be very different principles lurking in biological systems of which we are currently unaware. These would not be apparent to us from the sophisticated arsenal of experimental techniques we now use to understand genetic regulation. These tools are primarily designed to tease out specific interactions relating a few components from the large number that are present in the genome. On the other hand, hypothetically, to observe collective regulation, one needs to be able to examine thousands of components simultaneously, each one having a minuscule effect, but collectively, they produce precisely controlled regulation. An analogy with artificial neural circuitry might make this point clearer. In pattern recognition systems, where one desires to classify different images, most of the neurons fire in response to essentially all images that are presented. A single neuron is involved in the recognition of hundreds of thousands of images. Yet by precisely controlled collective interactions between units, very specific and accurate classification is achieved. Even in the case where all neurons can be probed simultaneously, it can be very difficult to understand how the circuitry operates, because the collective interaction of many components is not conducive to the kinds of explanations used in more normal digital circuitry. The same is expected for biological neural circuits as well. GWAS gives some insight but even with the massive amount of data being collected, the genomic circuitry is still greatly undetermined.

2. Relation to Boltzmann Machines

The purpose of a Boltzmann machine [13, 25, 26] is to learn a set of input/output pairs, and generalize from that information. If a set of inputs is presented, a corresponding set of outputs should be retrieved. This is accomplished as follows.

Consider a set of variables, often referred to as "spins", s_i that can take on only the values ± 1, but can change their values over time, as the system

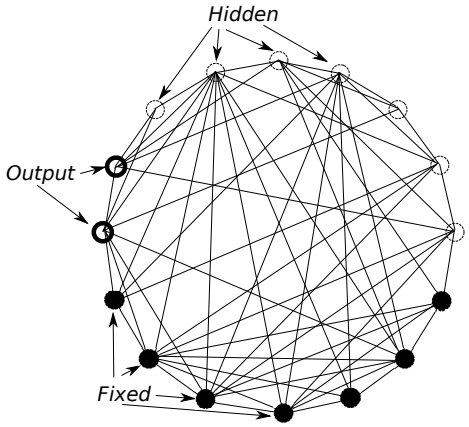


Figure 1: Illustration of a Boltzmann machine network. The black circles and unfilled thick circles are inputs and outputs of the network, respectively, and are fixed in value while the system is trained by an updating procedure. This results in output units (in this case there are two), that are trained so that their values depend on the values of the input neurons. The remaining dashed unfilled circles are hidden units that are used to process the input units.

is updated, for example using the Metropolis Monte Carlo algorithm at some finite temperature.

They interact via a connectivity matrix J_{ij} that couples spins i and j. The couplings are chosen by a method outlined below, to optimally give the correct outputs. To do this updating, one writes down an energy function, or Hamiltonian, for this system

$$H = -\sum_{i=1,j=1}^{N} J_{ij} s_i s_j \tag{1}$$

Out of these N spins, there are N_f spins that can sometimes be fixed. This is sometimes referred to as the "clamping" phase. These N_f spins are often referred to as "visible" units. When they are fixed, they have constant values during the updating procedure. These represent the training set of data for the machine. One can think of their effect as externally imposing constraints on the dynamics of the other $N_v \equiv N - N_f$ variable spins. Thus the first N_v spins $s_1, \ldots, s_{N_v}$ can vary and the last N_f spins $s_{N_v+1}, \ldots, s_N$ are, sometimes, frozen. Of the last N_f spins, one can regard N_i as inputs and $N_o = N_f - N_i$, as outputs. These are the input/output pairs mentioned above. This is depicted in Fig. 1, where the black circles represent the input units (that is spins), and the outputs are the black unfilled circles. The dashed unfilled circles represent the hidden units. The interactions J_{ij} are depicted by straight lines connecting the spins.

The algorithm starts by fixing these N_f spins to one of the input/output pairs while the system goes through many updating cycles. Then the system is allowed to run with the N_f visible spins now being *unclamped* so that they

are no longer fixed and are updated in the same way as the rest of the spins. Comparing the statistics of the unclamped and clamped simulation allows one to evolve the weights J_{ij}, moving the system closer towards the optimal set of couplings. As this happens, the temperature of the system is also slowly decreased. This is an application of simulated annealing [27].

The next step in this process is to change the fixed spins to another input/output pair and the above unclamping and annealing process repeats. Eventually, the connectivity matrix will have evolved to one that results in a system that has optimally learned these input/output pairs. That is, if one of the input sets is presented, it responds by giving the corresponding output.

The Hopfield model [14, 15, 16, 17, 18, 19] considers the case of no hidden units. In that case, there are M input/output pairs to be learned. There is no logical distinction here between inputs and outputs, and any subset of the spins could be presented, with the expectation that the rest would correctly flip to the desired outputs. Let us suppose that we want to learn M separate spin configurations. Let us denote the αth spin configuration by $\{t_1^\alpha, \ldots, t_N^\alpha\}$, for $\alpha = 1, \ldots, M$. Then a Hebbian rule can be used to explicitly write down the couplings without going through any simulated annealing procedure,

$$J_{ij} = \mathcal{N} \sum_\alpha t_i^\alpha t_j^\alpha \tag{2}$$

with a normalization factor $\mathcal{N}$, that varies depending on the author, and will be chosen later on in what follows. The number of patterns that can be reliably stored is proportional to N [18] but this will also depend on the correlations between the patterns. The mean field solution of these equations at inverse temperature β is

$$s_i = \tanh(\frac{\beta}{N} \sum_j J_{ij} s_j) \; i = 1, \ldots, N. \tag{3}$$

With the Hebbian couplings of Eq. 2, and statistically independent t_i^α's (with mean zero), the choice $s_i = t_i^\alpha$ can be shown [12] to satisfy these equations.

One can also use the same couplings in a Boltzmann machine with hidden units. This will not be optimized over the choice of spin values for the hidden spins, but for this choice of coupling, it will lead to a recall of all the visible units.

3. The System

RNA is used in an enormous number of ways in biology. The majority of transcripts in human cells do not directly code for proteins, but are non-coding [4] with functions that are mostly not well understood. Here I make the conjecture that that much of this RNA is involved in the kind of collective regulation described above.

The system studied here is a collection of N RNA species that interact through base pair binding and unbinding and other possible weak associations.

Each molecule can bind to itself and other molecules. RNA-RNA interactions in physiological conditions allow for the formation of secondary structure, and therefore must also allow for the binding to different molecules, as such interactions are of identical strength and mathematical form. The amount of non-coding RNA (ncRNA) in a cell is quite substantial [28]. One therefore expects that these RNA molecules have a considerable degree of interaction with each other, and furthermore, that they will bind frequently to other molecules such as some proteins. This is discussed further in section 6.1.

The inputs to the cell, such as signaling molecules, are well known to affect the transcription of DNA to RNA, and in particular, messenger RNA (mRNA). Because of the substantial interaction between RNA molecules, this in turn will affect the concentrations of all of the RNA. Some of these other RNA molecules will be involved with protein translation. By promoting or suppressing protein translation, these RNA concentrations will affect the function of the cell. Thus the cell inputs are "processed" by a complex system of DNA, proteins, and RNA, to produce or modify cell outputs. This paper investigates the possibility that it is the substantial and complex interactions between the RNA molecules that underlie the sophisticated computations used by the cell in determining how it will respond to different inputs.

The basic linking between inputs and outputs, is similar to the standard regulatory mechanisms involving binding of regulatory proteins along different sites on the genome [29]. In contrast with the RNA proposal above, such bindings require much more specificity, in much the same way as a conventional digital circuit requires precise connections between its elements. A system of promiscuously binding RNA molecules in all likelihood, is incapable of such specificity, and instead must resort to collective behavior, in analogy with artificial neural networks.

There are two basic components that are needed here to achieve such a collective computation. The first is that there is a chemical equilibrium between N molecular species undergoing reversible reactions. Such an equilibrium is well understood [30]. One assumes, with some justification given in Sec. 6, that because of the relatively high concentration of molecules, the system can quickly reach equilibrium concentrations. The steady state concentrations of different species is what will be of interest here.

On the other hand, the lifetime of RNA molecules is finite, and this means that in steady state, they require creation. The way that this happens is crucial to the second component to this model and is discussed in detail below. This is a subtle problem and certainly requires experimental verification. In order for our set of RNA molecules to perform sophisticated computations, I looked for a simple rule that would be biochemically plausible, and give behavior analogous to Boltzmann machines. As we will see, this can be achieved if the creation rate of a species depends on the ratio of bound to unbound RNA. Different biological scenarios for accomplishing this will be discussed later in Sec. 6.2.

In the following sections, it will be shown that these two rather simple assumptions involving RNA equilibration and creation, can perform collective computations. We first will consider an intermediate model where the mathe-

matical relationship with learning algorithms is the most apparent. I will then show how this can be simplified further to come up with a more biologically plausible mechanism.

3.1. Chemical equilibration

We consider N different chemical species of RNA that bind and unbind at rates that depend on their primary sequences. For example, complementary sequences will be most strongly bound. For the moment, assume that there are fixed total concentrations of each species $C_1, C_2, \ldots, C_N$. The corresponding unbound concentrations are denoted as $\rho_1, \rho_2, \ldots, \rho_N$. For simplicity I will assume binary reactions between molecules i and j,

$$i + j \rightleftharpoons ij, \tag{4}$$

and that there are no higher order reactions present. Including more complex reactions should not preclude the scenario presented here from working, and is an interesting topic for further investigation. We will denote the concentration of two molecules i and j, that are bound together, by ρ_{ij}.

The equilibrium constant [30] for such a reaction is $K_{ij} = \rho_{ij}/\rho_i\rho_j$. This implies [7]

$$\rho_i = \frac{C_i}{1 + \sum_j \rho_j K_{ij}} \tag{5}$$

for $i = 1, \ldots, N$.

In this model, the set of equilibrium constants K_{ij}, is fixed during the lifetime of a cell, as would be expected. It is posited that these evolve through mutation, to be able to take on arbitrary values, within some physical limits. Because binding between two different molecules will take place preferentially along certain species specific regions, there are enough degrees of freedom for these binding affinities to be chosen independently. Even if there is some dependence, there are many possible choices for couplings that lead to useful learning.

Note also that the system can be rescaled by a factor C_0 with units of density, by defining rescaled primed variables $K_{ij} = K'_{ij}/C_0$, $\rho_i = \rho'_i C_0$, $C_i = C'_i C_0$ and Eq. 5 will remain the same in the rescaled primed variables. This allows us to suitably rescale the equilibrium constants.

3.2. Creation of new RNA

As mentioned above, degraded RNA molecules must be replaced by new ones, and the most subtle part of the mechanism proposed here is how molecule production is regulated. In this model, one has two requirements related to RNA creation. First, that the rate will be controlled by the fraction of total to unbound molecules of the same species. Second, we require a mechanism to regulate the total concentration of RNA molecules. This latter requirement is fairly uncontroversial, and this can be regarded as a homeostatic feedback mechanism. More specifically, the first requires that the generation of a particular RNA species increases as the ratio of total to unbound RNA increases. I

will discuss how these mechanisms could operate in more detail in Sec. 6.2 and will now briefly summarize the two main scenarios that could give rise to this kind of regulation.

As mentioned in the introduction, RNA-dependent RNA polymerase (RdRP), can directly copy RNA molecules without the presence of DNA. For a given species, we require that the rate of copying will depend inversely on the amount of free RNA present.

A more likely explanation for how such a dependence could be achieved is through DNA cis-regulatory elements, regulating RNA transcription. There are already similar kinds of regulation that have been observed [24], and this will be discussed further in Sec. 6.2. However this kind of regulation is of course not evidence for the existence of the kind of computation proposed here, but suggests that this kind of mechanism would be worthwhile to look for experimentally.

In the first model that is considered below, the transcription rate will depend on other factors as well, (see Sec. 3.5), but it is shown later in Sec. 5, that one can dispense with these additional dependencies and produce a model with a transcription rate as described above, making this much more biologically plausible.

The point of this model is as a proof-of-concept. In addition, there are many variants, some of which have already been mentioned, such as higher order interactions, that could also perform collective computation. The main point of the following analysis is to make the case that this sort of mechanism is plausible.

3.3. Dynamics of concentration and transcription rates

Assuming no degradation or creation of RNA, the system will go to equilibrium concentrations given by Eq. 5, which determines the unbound RNA concentrations given the total concentrations of all the RNA molecules. However because of degradation and creation, the actual concentrations will differ from the equilibrium case.

In the framework described here, certain RNA species, for example mRNA, will act as inputs and for simplicity, these inputs will be assumed to have fixed values, while the remaining species will have time variation in their bound and unbound concentrations, due to RNA degradation, creation, and the interactions with other molecules. Out of the N species of RNA, one can say the N_v of the concentrations can vary and N_f of them are are fixed, with $N_v + N_f = N$.

If a closed system is not initially in equilibrium, the unbound concentrations will vary in time, asymptotically approaching the equilibrium values. The actual dynamics will be very complex and there will be a spectrum of relaxation times associated with the RNA concentrations' dynamics.

Assuming that binding and unbinding takes place on a timescale much shorter than the lifetime of an RNA molecule, then this allows us to describe dynamics with only one relaxation time τ_ρ through a standard first order kinetic equation.

$$\tau_\rho \frac{d\rho_i}{dt} = -\rho_i + \frac{C_i}{1 + \sum_j \rho_j K_{ij}} \tag{6}$$

for $i = 1, \ldots, N$. The relaxation timescale is assumed to be very short compared to the other processes described below, and therefore their detailed relaxation spectrum on a longer timescale is unimportant. This will be justified further in Sec. 6.1.

We are now in a position to quantify the mechanism of RNA creation discussed above. For simplicity, this model assumes a degradation timescale τ_C that is independent of molecular species. The rate of transcription of the *ith* species is regulated by a process that depends on both the total concentration C_i of a species, and all of the unbound ρ's.

$$\tau_C \frac{dC_i}{dt} = -C_i + f(C_i, \{\rho_k\}) \tag{7}$$

for $i = 1, \ldots, N_v$. Later I will show how this dependence can be considerably simplified to make it more biologically plausible. One expects that the process of degradation and production takes place at a much slower time scale than the molecular equilibration mentioned above, so that $\tau_C \gg \tau_\rho$. This is born out by estimates using empirical data, as discussed in Sec. 6.1. The function f gives the rate at which molecules of type i are being created through the kind of mechanism described above in Sec. 3.2 and in Sec. 6.2.

The remaining C_i, $i = N_v + 1, \ldots, N$, will act as inputs as described above, and those concentrations will not vary in time. Processes external to the ones considered here, are maintaining those levels; for example, the transcription of mRNA molecules that are acting as fixed inputs.

3.4. Connection to Boltzmann Machines

It is useful to connect the machine learning system discussed in Sec. 2 to the genomic system above. The variables of interest for the Boltzmann Machine are the spin variables $s_1, \ldots, s_N$. One would like to relate these to the concentration of unbound RNA $\rho_1, \ldots, \rho_N$. In this case, the s_i will no longer only take on the values ± 1, but can vary over the reals. I chose a linear relation between the two sets of variables

$$\rho_i = \delta \frac{1 + s_i}{2} + b \tag{8}$$

where δ and b are constants. From the form of solution in Eq. 3 the s_i still must be bounded by ± 1, and therefore ρ_i is bounded between b and $b + \delta$. These bounds are chosen to be biologically sensible, meaning that the unbound concentrations, $\{\rho_i\}$, need not become arbitrarily small or large for the mechanism proposed here to work.

Corresponding to the learned patterns t_i^α in Eq. 2, will be the *learned unbound concentrations* defined as

$$p_i^\alpha = \delta \frac{1 + t_i^\alpha}{2} + b \tag{9}$$

In analogy with learning algorithms, M patterns of ρ are being stored, with the αth pattern having unbound concentrations of $\{p_i^\alpha\}_i$.

Similarly, one would like to relate the Boltzmann Machine couplings in Sec. 2 to the equilibrium constants K_{ij} above, through

$$K_{ij} = \epsilon \frac{1 + J_{ij}}{2} + a \tag{10}$$

where a and ϵ are both constants. If one places the restriction $|J_{ij}| \leq 1$ for all i and j, by appropriate normalization of Eq. 2, this means that $a < K_{ij} < a + \epsilon$. This allows us to choose physically sensible values for the equilibrium constants.

3.5. Creation rate

To relate the Hopfield model solution Eq. 3 to the kinetic equations for our RNA system, it is necessary to choose a specific and seemingly complicated form for f in Eq. 7. We are interested in the system's steady state behavior, where all time derivatives are zero, which simplify Eqs. 6 and 7. This requires that the choice for f be

$$f(C_i, \{\rho_k\}) = \frac{C_i}{\rho_i} S(\frac{4}{\epsilon}(\frac{C_i}{\rho_i} - 1) - 2(1 + 2\frac{a}{\epsilon}) \sum_{j=1}^{N} \rho_j - \frac{2(\delta + 2b)}{\epsilon} \sum_j K_{ij} + (\frac{2a}{\epsilon} + 1)(\delta + 2b)N) \tag{11}$$

for $i = 1, \ldots, N$. We will see in the section, why this is.

The function $S(x)$ is chosen to be

$$S(x) = \frac{\delta}{2}[1 + \tanh(\beta x/N)] + b \tag{12}$$

But there is a large class of sigmoidally shaped functions that would also work.

Eqs. 6, 7, 11, and 12 along with arbitrary initial conditions, fully define the dynamics of the unbound and bound RNA concentrations, given a set of equilibrium constants K_{ij}. The next section explains the equivalence with learning algorithms in more detail.

4. Equivalence of RNA system to machine learning algorithm

We are interested in the long time steady state solution, where there is no time dependence and therefore all time derivatives are zero.

Eq. 6 becomes 5 in this limit, and therefore

$$\sum_j K_{ij} \rho_j = \frac{C_i}{\rho_i} - 1. \tag{13}$$

Similarly, Eq. 7 becomes

$$C_i = f(C_i, \{\rho_k\}). \tag{14}$$

Substituting in Eq. 11

$$C_i = \frac{C_i}{\rho_i} S(\frac{4}{\epsilon}(\frac{C_i}{\rho_i} - 1) - 2(1 + 2\frac{a}{\epsilon}) \sum_{j=1}^{N} \rho_j - \frac{2(\delta + 2b)}{\epsilon} \sum_j K_{ij} + (\frac{2a}{\epsilon} + 1)(\delta + 2b)N) \tag{15}$$

Canceling the C_i's and using Eqs. 13, 8 and 10, solving for s_i, and substituting Eq. 12 finally gives the same form as Eq. 3,

$$s_i = \tanh(\frac{\beta\delta}{N} \sum_j J_{ij} s_j) \tag{16}$$

It is not necessary that a tanh function be used here. A variety of sigmoidally shaped curves should have the correct properties, with similar efficacy.

The above analysis does not show that these equation will lead to this steady state solution, and indeed, if the time scales are not as described here, it can lead to different steady state behavior. Next, I will explore this problem numerically to find out if the equivalence to the above solution is viable, and if it does lead to machine learning.

4.1. Numerical results

The above model was implemented numerically. The system had $N = 50$ RNA species with unbound concentrations $\{\rho_i\}$ and total concentrations $\{C_i\}$ as described above, and these evolve over time according to the Eqs. 6, 7, 11, and 12. The equilibrium constants K_{ij}'s were chosen according to Eq. 2 where we analyzed the retrieval of $M = 3$ patterns. The t_i^α were chosen randomly to be ± 1 and these correspond to values of ρ given in Eq. 9. After evolution for sufficient time to have converged, the program checked to see if the pattern of ρ's found was one of the three patterns that were encoded in the K_{ij}'s.

When the transformation of Eq. 10 was applied, the final K's were scaled so that their values were between a and $a + \epsilon$. The values of the ρ's were also rescaled according to Eq. 8. It appeared that the values of these rescaling parameters, a, b, ϵ and δ, did not have a strong effect on convergence of the model.

The ratio of the two timescales τ_C/τ_ρ needed to be sufficiently large to obtain consistent convergence over a wide range of initial conditions. A ratio of $\tau_C/\tau_\rho = 100$ was found to work in all cases. With smaller values, such as $\tau_C/\tau_\rho = 10$, convergence worked well for some initial conditions but not for all of them.

When starting with arbitrary initial ρ_i's, the corresponding values of C_i were chosen by rearranging Eq. 13

$$C_i = \rho_i(\sum_j K_{ij}\rho_j + 1) \tag{17}$$

The equations were evolved using an explicit embedded Runge-Kutta-Fehlberg 4(5) method, with a step size of 0.1.

Several important properties of the system's dynamics were studied, the basin of attraction starting from ρ's that were different from the initial patterns. Another question considered, was how altering the optimal K_{ij}'s influenced the final patterns found. And finally, how the number of hidden units influence the system's performance. The following two subsections study sensitivity to deviations in the ρ's and deviations in the K's. The third subsection considers the effects of clamping some of the concentrations to fixed values, to study how well such systems perform as Boltzmann machines.

4.1.1. Sensitivity to unbound concentrations

The first numerical study tested out the basin of attraction of initial values of the ρ_i's. Because the ρ's continuously vary in time, to compare the converged solutions to the binary patterns t_i^α, the ρ's were partitioned so that they corresponded to -1 if $\rho_i < b + \delta/2$, at $+1$ otherwise, which is seen from the mapping between the two systems in Eq. 9.

An initial pattern was altered from t_i^α so that it differed randomly at n locations. That is, the Hamming distance was set to n. Then the system was evolved from this condition to see if it would relax back to that same pattern $\{t_i^\alpha\}_i$. Note that because of symmetry, both $\{t_i^\alpha\}_i$ and $-\{t_i^\alpha\}_i$ are possible solutions that should be considered when comparing for convergence. Any deviation from the trained pattern was considered a mistake.

At every initial Hamming distance n, 10 independent sets of M patterns were generated. For each of these sets, the program started with 10 randomly altered patterns that were evolved for each of the M patterns. Altogether, this represents 300 separate runs for each Hamming distance studied.

I also investigated making two separate kinds of alterations to the initial conditions: ones that conserve the total number of 1's and -1 relative to the learnt pattern $\{t_i^\alpha\}_i$, and ones that allow this total to vary. This becomes an important distinction later on, when a more universal version of this model is considered.

In Fig. 2, and most subsequent plots, two sets of values were used, in (a), $a = b = 0.4$, and $\epsilon = \delta = 0.6$ and in (b) $a = b = 0.001$ and $\epsilon = \delta = 0.999$ (see Eqs. 8 and 10). As will be seen, for most quantities, the results are quite insensitive to these choices.

The fraction of mistakes as a function of the Hamming distance cutoff is shown in Fig. 2(a) and (c). The three lines in (a) and in (c), show the results for different values of β, $\beta\delta/N = 0.5$, 2, and 8. With less accurate iteration methods, it was found that $\beta\delta/N < 8$ was not stable. However with this Runge Kutta method, the differences between the results are fairly minor. The precise sigmoidal shape $S(x)$ in Eq. 12 is clearly not important.

I now consider the second kind of initial conditions mentioned above, where the initial condition does not change the total unbound concentration. To generate such initial conditions, the program starts with the target learned unbound concentrations $\{\rho_i\} = \{t_i^\alpha\}_i$ and makes alterations to the $\{\rho_i\}$ in random pairs i, j, so that the sum of the $\rho_i + \rho_j$ stays constant. The results are shown in Figs. 2(b) and (d).

4.1.2. Sensitivity to equilibrium constants

I now consider the effect that changing the K_{ij}'s has on the patterns that are retrieved. With the usual Hopfield model, it is known to be quite robust to changes of the connectivity strength, which greatly contrasts with usual digital architecture. We will now see to what extent this still carries over here.

The algorithm picks i's and j's at random and mutates them, in a way which is equivalent to $J_{ij} \to -J_{ij}$. Taking into the account that the parameters we chose have $a + \epsilon = 1$, this is equivalent to taking $K_{ij} \to 1 - K_{ij}$. The average Hamming distance $H(\{t_i^\alpha\}_i, \{\rho_i\}_i)$ is computed by translating the ρ's into corresponding discrete spin variables. This way, we are counting the number of differences between the t_i^α's and the final pattern, $s_{final,i}$ that emerge. The Hamming distance is normalized by dividing by N, that is $h = H(\{t_i^\alpha\}, \{s_{final,i}\})/N$. This is computed as a function of the number of mutations n_K made to the K_{ij}'s. To normalize this, we define

$$f_K \equiv \frac{2n_K}{N(N-1)}. \tag{18}$$

The program performs these kinds of mutations with all other parameters identical to the ones used above, and the results are shown in Fig. 3.

4.1.3. Clamping input concentrations

Now consider clamping N_f of the C_i's so that they are fixed to predetermined values as the biochemical network evolves in time, to see if it can correctly associate those clamped inputs to outputs. This is the kind of task that is performed by a Boltzmann machine. Fig. 1 illustrates this process. The filled black circles represent the clamped inputs, and are coupled through the K_{ij}'s, represented by lines, to all other units. A single unit, i, represents the concentrations ρ_i and C_i. The unfilled circles have ρ_i's and C_i's that vary, and two of these circles represent outputs. To test out this capability, the program was initialized as before, with $N = 50$ and choosing $M = 3$ separate random patterns $S_i = t_i^\alpha$ which are then translated into ρ_i's, again using Eq. 9. The couplings K_{ij} are chosen as before as well.

N_f of the C_i's, for $i = N_v + 1, \ldots, N$, were fixed, and the other N_v units were unclamped as before, as described by Eq. 7. The initial conditions were varied, by randomly scrambling the remaining values of ρ_i, $i = 1, \ldots, N_v$. As usual, the corresponding initial C_i's were chosen through Eq. 17.

Out of the $N_v = N - N_f$ ρ_i's that vary, one can regard two of these as output units and the rest as hidden. We would like to know how well, given the fixed inputs, the system evolves to finally recall these two output units.

Fig. 4 shows the fractional number of mistakes plotted versus the number of variable units, N_v, for two different temperatures, $\beta\delta/N = 0.5$ and 4.0.

5. More universal regulation

The numerical results of the previous section illustrate that for a wide range of parameters, this genetic biochemical network has capabilities quite similar to

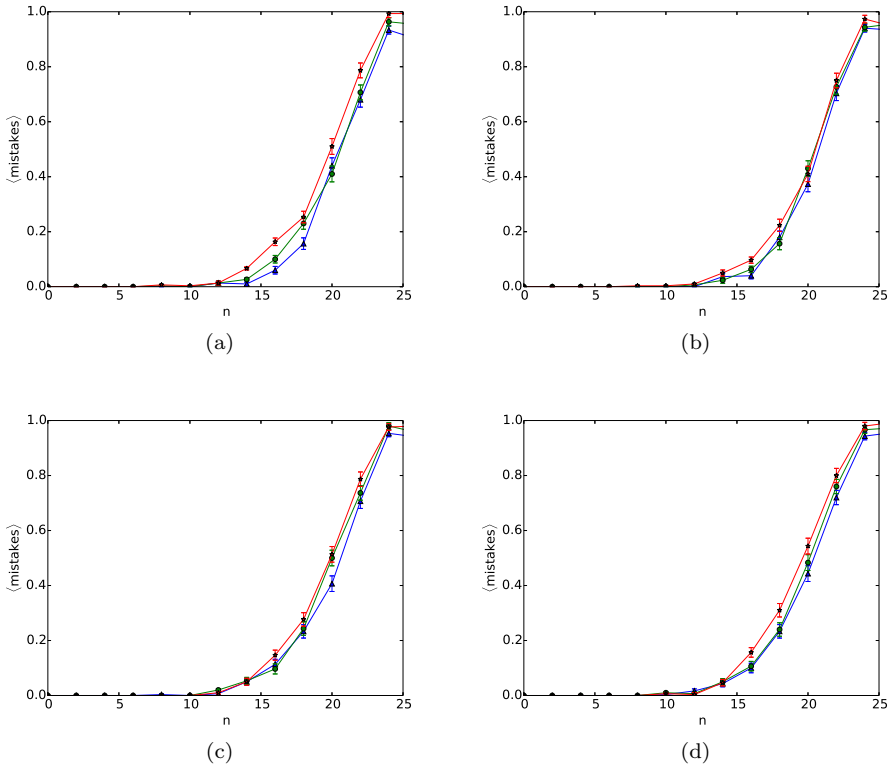
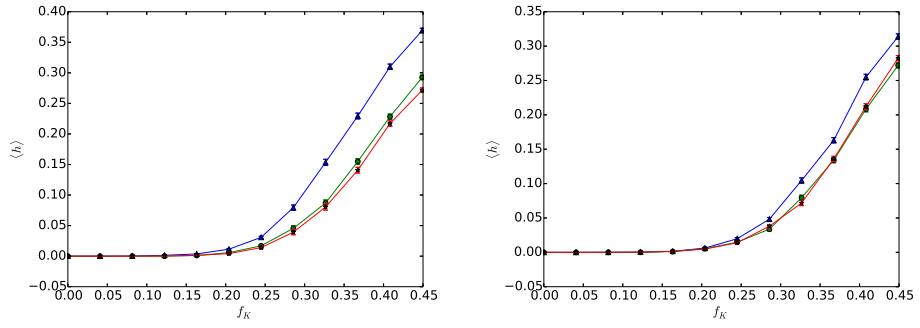


Figure 2: (a) The average fractional number of mistakes made as a function of the Hamming distance, n, between the initial state and the pattern. Here the total number of RNA species is $N = 50$ and the total number of patterns to be recalled is $M = 3$. $a = b = 0.4$, and $\epsilon = \delta = 0.6$ (see Eqs. 8 and 10). The triangles have $\beta\delta/N = 0.5$ the circles, $\beta\delta/N = 2$, and the stars show $\beta\delta/N = 8$. The lines are simply a guide for the eye. (b) The same case as in (a) except that the total unbound concentration of the initial state is unchanged. (c) and (d) are the same as (a) and (b) respectively, but with $a = b = 0.001$, and $\epsilon = \delta = 0.999$

those of powerful machine learning algorithms. The main criticism of this system is the rather contrived nature of the function f in Eq. 11. This complicated form was designed to give the same steady state solutions as the analogous machine learning system. But it is not clear how this could be implemented biologically. I now show how this mechanism can be greatly simplified, leading to a much stronger case for biological relevance.

Let us start by writing Eq. 11 as

$$f(C_i, \{\rho_k\}) = \frac{C_i}{\rho_i} S(\frac{4}{\epsilon}(\frac{C_i}{\rho_i} - 1) - \mathcal{A}) \tag{19}$$



(a) (b)

Figure 3: (a) The average normalized Hamming distance is plotted versus fractional mutation frequency of equilibrium constants f_k. Triangles correspond to $\beta\delta/N = 0.5$, circles $\beta\delta/N = 2.0$, and stars $\beta\delta/N = 4.0$. $a = b = 0.4$, and $\epsilon = \delta = 0.6$ (see Eqs. 8 and 10). (b) The same except that $a = b = 0.001$, and $\epsilon = \delta = 0.999$

with

$$\mathcal{A} = 2(1 + 2\frac{a}{\epsilon})\sum_{j=1}^{N}\rho_j + \frac{2(\delta + 2b)}{\epsilon}\sum_j K_{ij} + (\frac{2a}{\epsilon} + 1)(\delta + 2b)N) \tag{20}$$

The complication here is that $\mathcal{A}$ is not a constant but depends on unbound densities ρ_i. However, it only depends on the sum of all of these. Consider initial conditions that still differ from the patterns p_i^α but have the same total sum. We ask if replacing $\sum \rho_i$ by a constant value will influence the steady state, that is, Eq. 16. Therefore it is possible to define a more "universal" creation function as follows

$$f(C_i, \{\rho_k\}) = \frac{C_i}{\rho_i} S(\frac{4}{\epsilon}(\frac{C_i}{\rho_i} - 1) - 2(1+2\frac{a}{\epsilon})\sum_{j=1}^{N} p_j^\alpha - \frac{2(\delta + 2b)}{\epsilon}\sum_j K_{ij} + (\frac{2a}{\epsilon} + 1)(\delta + 2b)N) \tag{21}$$

where the sum of the ρ_i's has been replaced by a sum over pattern α, p_i^α. One can now follow the same steps as were employed in Sec. 4 to relate this biochemical system to the machine learning spin system. In this case, Eq. 15 now has the term in the argument of the function S, $2(1+2\frac{a}{\epsilon})\sum_{j=1}^{N}\rho_j$, replaced by $2(1 + 2\frac{a}{\epsilon})\sum_{j=1}^{N} p_j^\alpha$. The argument of S now differs from its previous value by

$$\Delta_h \equiv 2(1 + 2\frac{a}{\epsilon})\sum_{j=1}^{N}(p_j^\alpha - \rho_j) = (1 + 2\frac{a}{\epsilon})\delta\sum_{j=1}^{N}(t_j^\alpha - s_j) \tag{22}$$

where in the last equality, Eqs. 9 and 8 have allowed us to translate this difference into spin variables. To determine the effect of this term, for simplicity,

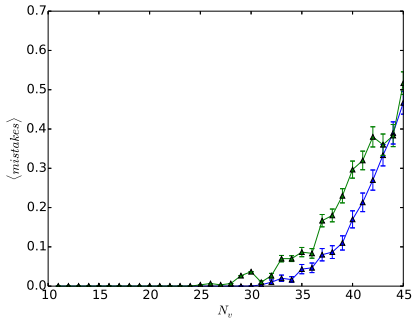


Figure 4: The average fractional number of mistakes made as a function of the number of variable units, N_v, in the Boltzmann machine analog illustrated in Fig. 1. The triangles are for the case $\beta\delta/N = 0.5$, and the stars are for $\beta\delta/N = 4.0$. Here $a = b = 0.4$, and $\epsilon = \delta = 0.6$ (see Eqs. 8 and 10).

consider the case where

$$\sum_i t_i^\alpha = 0 \ for \ \alpha = 1, \ldots, M. \tag{23}$$

Because these patterns are chosen at random, then for large N, this is a reasonable assumption.

Let us start by reviewing the simplest way that the mean field solutions in Eq. 3 or equivalently Eq. 16 can be understood[12]. This is done by utilizing the random statistical nature of the t_i^α's. Examining the argument on the right hand side of this equation, and using Eq. 2, we have that

$$\sum_{i=1}^{N} J_{ij} s_i = \mathcal{N} \sum_{i=1}^{N} \sum_{\beta=1}^{M} t_i^\beta t_j^\beta s_i. \tag{24}$$

Previously we chose $|J_{ij}| \leq 1$. Because the patterns are independent, to achieve this, the normalization factor $\mathcal{N}$, will be of order $1/\sqrt{M}$ (with additional logarithmic corrections that are not important to our conclusions). If I now change the "gauge", writing $\sigma_i \equiv s_i t_i^\alpha$, then

$$\sum_{i=1}^{N} J_{ij} s_i = \mathcal{N}[\sum_{i=1}^{N} t_j^\alpha \sigma_i + \sum_{\beta \neq \alpha}^{M} t_j^\beta \sum_{i=1}^{N} t_i^\beta t_i^\alpha \sigma_i] = \mathcal{N}[N t_j^\alpha \sigma_i + \sum_{\beta \neq \alpha}^{M} t_j^\beta \sum_{i=1}^{N} t_i^\beta t_i^\alpha \sigma_i] \tag{25}$$

If we choose the pattern $s_i = t_i^\alpha$, then $\sigma_i = 1$ for all i. The first term in the last equality is $(\mathcal{N}N) t_j^\alpha$. Because the patterns are random, the second term has a term of order $\pm\mathcal{N}\sqrt{NM}$, which for $N \gg M$, is negligible compared to the first term. In this limit, all of the other patterns $\beta \neq \alpha$ can be ignored, and this yields the mean field equation for a single pattern, the so-called "Mattis model" [18].

As is evident from Eq. 16, this is satisfied, and using the normalization factor $\mathcal{N} \propto 1/\sqrt{M}$, the size of the dominant term is of order $N/\sqrt{M}$.

Now let us return to the corrections to Eq. 16 given by Eq. 22. Doing the same kind of estimation, Δ_h has magnitude $\pm\sqrt{N}$. Therefore comparing the factors of N and M with Eq. 25, which is of order $N/\sqrt{M}$, Δ_h is negligible for $N \gg M$. $N \gg M$ is the case that we are already considering.

Therefore for the random patterns (typically used in machine learning problems such as the Hopfield model) with a constant sum, and for large N, replacing the summation of the ρ's by Eq. 21 is not expected to alter the steady state solutions of Eq. 16.

I will therefore only consider learnt patterns with the property that $\sum_i p_i^\alpha$ is a constant and does not depend on α. In the most important case[1] considered above, this corresponds to a machine learning problem satisfying Eq. 23. We therefore can write

$$P_{tot} \equiv \sum_i p_i^\alpha = N(\frac{\delta}{2} + b) \tag{26}$$

where the last equality used Eq. 8.

This shows that the creation mechanism can be modified so that it only depends on the ratio $\frac{C_i}{\rho_i}$

$$f(C_i, \{\rho_k\}) = \frac{C_i}{\rho_i} S(\frac{4}{\epsilon}(\frac{C_i}{\rho_i} - 1) - A) \tag{27}$$

where A maintains a constant value in time and only depends on the fixed parameters, such as the equilibrium constants K_{ij} and the total sum of learned pattern concentration P_{tot},

$$A = 2(1 + 2\frac{a}{\epsilon})P_{tot} + \frac{2(\delta + 2b)}{\epsilon} \sum_j K_{ij} + (\frac{2a}{\epsilon} + 1)(\delta + 2b)N) \tag{28}$$

Note that for the K_{ij} obtained using Eqs. 2 and 10 and 23, A does not depend on i. A more general treatment would include a dependence on i.

This approach assumes that $\sum_i \rho_i$ starts close to P_{tot}. If it does not, then an additional regulatory mechanism is needed to drive this sum towards P_{tot}. This would operate in a similar way to other homeostatic mechanisms. If P_{tot} deviates, the total RNA concentration should vary as well. Mechanisms would need to ensure that this stays at a well defined value. But even if we completely ignore such a general mechanism, it is straightforward to study this more universal model numerically and compare it to the results found earlier in Sec. 4.1. We will see that it still works surprisingly well.

Eqs. 7, 27 and 6 define the system of equations to be evolved in time. The creation of RNA is now much simpler to describe. It depends on the ratio of total concentration to unbound concentration.

This model will now be investigated numerically.

[1] Analysis of the Hopfield model for nonzero $\sum t_i^\alpha$ can also be performed [19].

5.1. Numerical results

The above model with this much simpler creation function, was studied using the same parameters as in Sec. 4.1, e.g. $M = 3$, and $N = 50$. As mentioned above, this assumes a general homeostatic mechanism, as considered earlier in Fig. 2(b) where the initial unbound concentrations preserve their total value, $\sum_i \rho_i$.

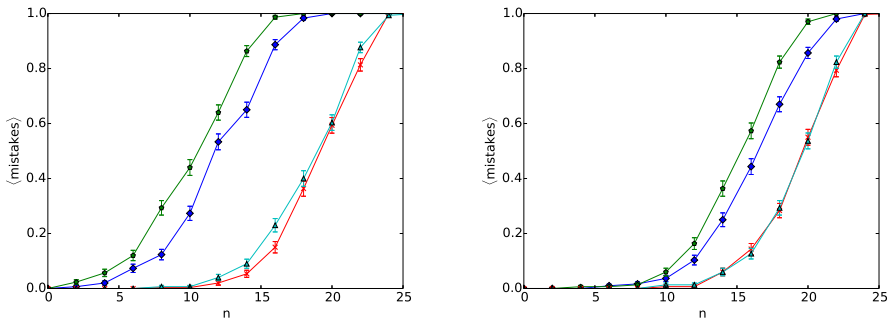
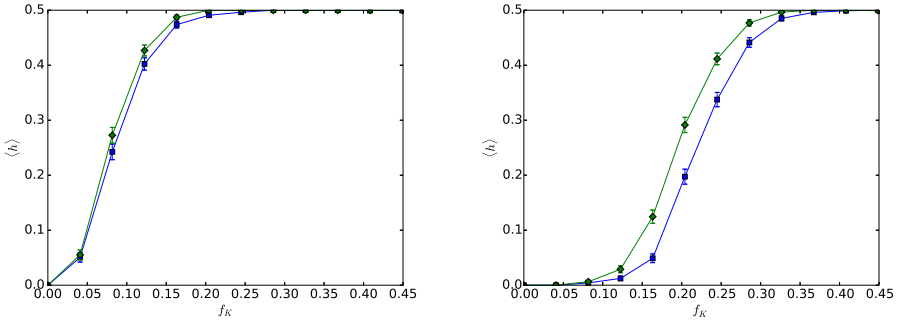


Figure 5: (a) The average fractional number of mistakes made as a function of the Hamming distance between the initial state and the pattern, n. Here the total number of RNA species is $N = 50$ and the total number of patterns to be recalled is $M = 3$. The triangles are the results for $\beta\delta/N = 4$ and the initial total unbound concentration equal to that of the patterns to be recalled. The pentagons are for the same parameters but the initial unbound concentration has one more up "spin". The crosses are with those same initial conditions but with $\beta\delta/N = 8$, and the diamonds are with one more up "spin". The lines are simply a guide for the eye. Here $a = b = 0.4$, and $\epsilon = \delta = 0.6$ (see Eqs. 8 and 10). (b) The same situation with $a = b = 0.001$, and $\epsilon = \delta = 0.999$.

To investigate how well this system works in more detail, I consider systems that are regulated so that the total concentration of unbound RNA starts off close to P_{tot}, but is otherwise scrambled. This was done with the same procedure as in Figs. 2(b) and (d). The graphs in both Figs. 5 (a) and (b), show the fractional number of mistakes as a function of the Hamming distance, n, between the initial state and the pattern to be recalled. The recall works best when the total unbound concentration is maintained at the correct final amount, and is less good when it deviates from that. This shows the necessity for carefully regulating the total RNA concentrations.

Similarly, the variation of the normalized Hamming distance as a function of the mutation frequency of the K_{ij}'s is shown in Fig. 6. In comparison with Fig. 3 it shows more sensitivity. This is not surprising, because the sum over the K_{ij}'s in Eq. 21 will no longer be the same, and this kind of variation was not taken into account in the analysis of the last section, only changes in the ρ_i's. Biochemical circuitry could be posited to further adjust A, but because mutations in the K_{ij}'s happen in the process of evolution, additional biochemical circuitry is not necessary if one allows for changes in the value of

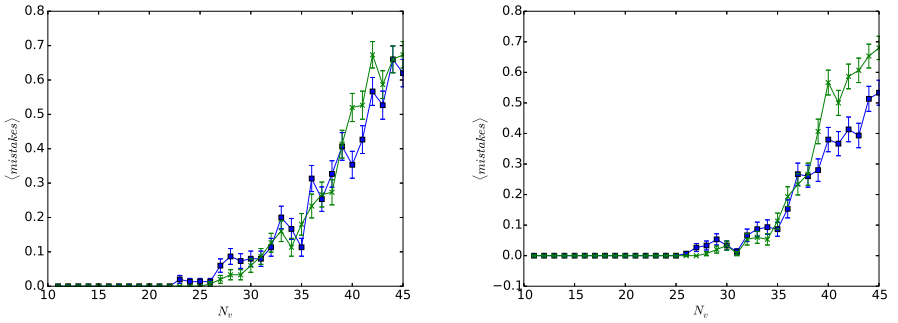
A to occur during evolution. Any detailed discussion on this topic becomes far too speculative to warrant serious consideration at this stage.



(a) (b)

Figure 6: (a) The average normalized Hamming distance is plotted versus fractional mutation frequency, f_K, of equilibrium constants K_{ij}. Diamonds correspond to $\beta\delta/N = 4.0$, and squares $\beta\delta/N = 8.0$. Here $a = b = 0.4$, and $\epsilon = \delta = 0.6$ (see Eqs. 8 and 10). (b) The same situation with $a = b = 0.001$, and $\epsilon = \delta = 0.999$.

In reference to Boltzmann machines, Fig. 7 shows the fractional number of mistakes plotted versus the number of variable units, N_v, for two different temperatures, $\beta\delta/N = 0.5$ and 4.0. It is quite similar to Fig. 4.



(a) (b)

Figure 7: (a) The average fractional number of mistakes made as a function of the number of variable units, N_v, in the Boltzmann machine analog illustrated in Fig. 1, here $a = b = 0.4$, and $\epsilon = \delta = 0.6$ (see Eqs. 8 and 10). The crosses are for the case $\beta\delta/N = 0.5$, and the squares are for $\beta\delta/N = 4.0$. (b) The same situation with $a = b = 0.001$, and $\epsilon = \delta = 0.999$

6. Discussion

6.1. Magnitude of RNA interactions

Now I estimate the density and timescales for non-coding RNA from the available data.

One expects that little or none of the long ncRNA in the cytoplasm will import back into the nucleus and therefore we can confine our attention to RNA that is preferentially localized to the nucleus. Given its regulatory role, it is not surprising that ncRNA is, on average, preferentially enriched in the nucleus, in contrast with mRNA that is exported into the cytoplasm. The ratio of nuclear to cytoplasmic ncRNA is approximately 1 [31]. In other words, approximately half of it is localized to the nucleus.

Given its predominantly regulatory function, it is not surprising that the total amount of ncRNA present in a cell is estimated to be lower than the total amount of mRNA. It is appears, on average that long ncRNA has approximately a tenth of the abundance of mRNA, although this number fluctuates substantially depending cell type, much more than for mRNA [28]. The total number of mRNAs in a mammalian cell is approximately [32] 5×10^5. This implies that the total amount of long ncRNA is approximately $N_L = 2 \times 10^4$ per cell nucleus.

A mammalian cell nucleus has a radius of approximately $r_n = 3\mu m$. This gives a long ncRNA density of $\rho_L = 3N_L/(4\pi r_n^3)$, or an average separation of $r_L = \rho_L^{-1/3}$, which is approximately $0.18\mu m$. Because of the heterogeneous nature of the nuclear environment, it is not easy to get a precise estimate for diffusion coefficients, but mRNA in the nucleus appears to have a diffusion coefficient $D \approx 0.1\mu m^2/s$ [33] which should be similar to that of ncRNA, although there will be a large range depending on the species. Therefore the time for an ncRNA molecule to move a distance r_L is on average $r_L^2/6D$, which is approximately, $.05s$. In that time, it will not necessarily encounter another ncRNA molecule, and this will increase the time scale by a factor of r_L/d, where d is a measure of the size of the molecule. For a 1000 base pair RNA, with a persistence of length of approximately 40 Å, this gives $d \approx 40nm$, and therefore $r_L/d \approx 5$. This means that the collision time is approximately $0.25s$.

The half-life for ncRNA in the nucleus is of order 30 minutes to an hour [6]. Therefore one expects that there is a large difference between the time scale for equilibration and for degradation of these ncRNA molecules, as assumed by the model here.

The free energy gain due to base pair conjugation is high, about 2 k_BT per base pair at biological temperatures [34]. This is why in general, RNA easily associates forming secondary structures. Some species of ncRNA might associate strongly with others, so that it binds irreversibly and is eventually degraded. Adding this possibility would complicate the analysis. Excess RNA from the more abundant species will be left unbound and can then bind with other RNA.

Another effect is the association of ncRNA with RNA binding proteins [35] (RBPs). By forming secondary and tertiary structure and associating with RBPs, the interactions between ncRNA is likely to be reduced. However in

this situation, one does not expect the ncRNA to be completely inert, but it seems plausible to suggest the existence of a large number of relatively weak attractive interactions between the ncRNAs that would be reversible. These kinds of interactions would then lead to the weak but promiscuous interactions required for collective regulation.

It should be pointed out that experimentally, duplexes between different RNA do form between ncRNA molecules [36]. In this work, the authors developed a sophisticated method to assess the presence of inter-RNA duplexes. It involved crosslinking existing RNA duplexes in vivo and digestion of the uncrosslinked RNA. Following a number of further steps involving ligation and uncrosslinking, the resultant RNA were processed utilizing high throughput sequencing. Using probabilistic modeling to suppress intrachain interactions, they obtained a large number of interactions between all major classes of ncRNA and mRNA, for example, small nuclear RNA (snRNA) and long intergenic noncoding RNA (lincRNA). Their method was able to reproduce many known interactions and studied, in detail, interactions between snRNA. The interactions were dominated by the most prevalent species, for example ribosomal RNA (rRNA), but there were interactions found between different lincRNA. There are clearly many interactions they found that could not be contributing to the mechanism proposed here, for example, rRNA, tRNA, and miRNA. But that still leaves a large fraction that cannot as yet, be ruled out. The main problem with the interactions found in this work [36], is that they are most likely of a type that would be too strong to be relevant to collective regulation, but it is possible that some of those would be weak enough to be reversible. Altogether, this work shows the abundance of strong duplex interactions between RNA-RNA interactions. The interactions hypothesized for the model here should be even weaker, and are therefore likely to be even more abundant. It is expected that these RNA molecules will also associate with proteins, and this will suppress duplex formation. This association with proteins will not necessarily preclude different RNA molecules from associating with each other, but will serve to weaken interactions between them.

Another work investigating RNA-RNA interactions for specific cases of ncRNA [37] was performed using RNA antisense purification. It would be interesting to pursue this kind of research further, to better quantify the abundance and strength of these ncRNA-ncRNA interactions experimentally. If indeed there are a large number of weak interactions with properties consistent with the collective regulatory mechanism propose here, it would make this possibility much more promising.

It should stressed that none of the above proves that the mechanism proposed is present in real biological organisms. But it does make the case for further investigation, to rule out if the weak interactions needed are too insignificant to give rise to collective regulation.

6.2. Mechanisms for creation

Above I described a biophysical mechanism capable of performing sophisticated computation, using RNA produced by the genome. The ability to make

high-level decisions based on its inputs has obvious advantages to a biological organism. Taking advantage of the large amounts of non-coding RNA produced by a cell, should increase the computational capabilities roughly according to the number of mutual interactions between the different RNA species. For example, if a mechanism similar to this was to be utilized, it would allow an organism to learn from its previous evolutionary history by encoding past environments in the values of the equilibrium constants $\{K_{ij}\}$. However, it is far from clear that a mechanism similar to what has been described here, does in fact exist.

In this section I give some potential ways that Eq. 27, which gives the rate of creation of a RNA species i, could be realized in practice. The crucial quantity that the system must measure is the fraction of unbound RNA. Fig. 8 plots the general form of this function. The specific parameters used here are $\delta = 0.4$ and $b = 0.6$. The curve starts at $C/\rho = 1$ and increases from there. This creation rate deviates subtly from linearity, and a large number of functional forms with this general shape would be suitable. For example we have already seen that β can be considerably altered with only a modest effect on performance. The choice of the tanh function was also not necessary and a wide variety of sigmoidal shaped functions are expected also work as has been investigated in neural network models [12]. I will now discuss what kinds of models could be expected to give this general shape for the creation rate.

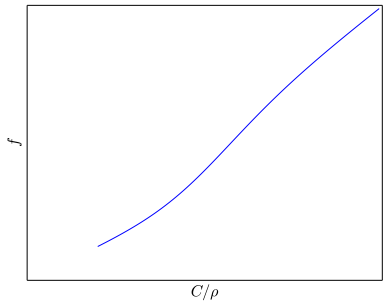


Figure 8: The creation rate of RNA given by Eq. 27 as a function of the total to unbound RNA C/ρ, for a given RNA species.

Let us first discuss a non-genomic mechanism, which is the most direct, but least likely to exist biologically. There is evidence, in humans [23], that there is a biochemical pathway recreating a similar function to RNA-dependent RNA polymerase (RdRP) [22]. RNA molecules could in principle be copied without reference to DNA. But in this case, the rate of transcription should depend on the ratio of C/ρ, that is, the total to unbound RNA of a single species. One mechanism to do this would be to have a double stranded RNA sensor. Toll-like [38] double stranded RNA sensors do exist, such as TLR3 [39] but these are membrane spanning however. This possibility is shown pictorial in Fig. 9(a). A less fanciful mechanism is illustrated in Fig. 9(b). Here the putative RdRP

is regulated by a site on it, shown as a circle. If RNA binds to this site, it will inhibit transcription. The most likely RNA to be bound is the same species that is being copied due to its close proximity to the RdRP. This potential binding process is shown by the arrow going from the end of the transcribed RNA, and pointing to the binding site. If a third RNA molecule, shown in light gray, associates with the copied RNA, it will inhibit binding to this site. This will give enhanced RNA creation as the ratio of total to unbound RNA increases.

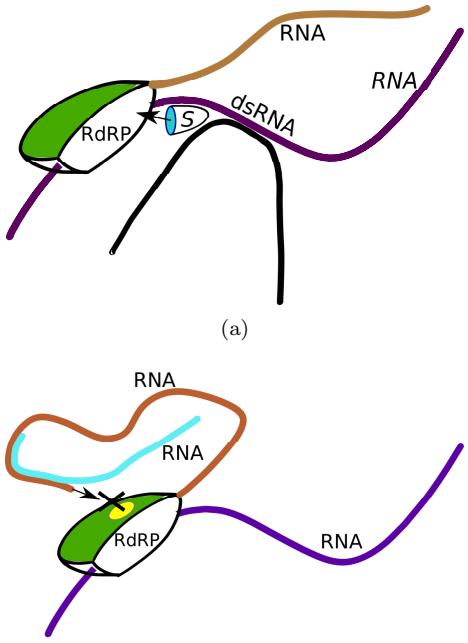


Figure 9: Two possible mechanisms for enhanced RNA transcription when additional RNA has been bound. (a) A RNA-dependent RNA polymerase (RdRP) copies an RNA molecule that is bound to another one forming a region of double stranded RNA (dsRNA). An RNA sensor, S, detects the dsRNA which then regulates the rate of transcription of the RdRP. (b) A site on RdRP, shown by the circle, represses transcription when RNA is bound to it. This is indicated by the arrow pointing from the RNA end to the binding site. The transcribed RNA is inhibited from associating with this regulatory site, by binding to another RNA molecule, shown in light gray.

Now consider more conventional and promising genomic mechanisms that give creation rates similar to Fig. 8. There are a number of theoretical possibilities for how the creation of RNA can depend on the ratio of total to unbound RNA. The unbound RNA can interfere [40] with the translation of an activator protein specific to the RNA species being transcribed from the DNA. The larger the amount of free RNA, the lower the rate of activator production, and hence the lower the rate of RNA production.

A more concrete possibility is a similar mechanism that is known to operate in some situations [41, 42, 24]. Many ncRNAs are transcribed around DNA

regulatory elements, such as enhancers. These ncRNA appear to increase the binding of transcription factors, which increases transcription rate. For example, in fission yeast *Schizosaccharomyces pombe*, transcription of ncRNA by RNA polymerase II (RNAP II), from the promoter region of $fbp1^+$ (fructose-1,6-bisphosphatase 1), has been shown to depend on the amount of that ncRNA that is present. The reason for this is due to the ability of this ncRNA to facilitate the binding of a transcription factor Atf1 on the $fbp1^+$ promoter [42]. The mechanism for this has been hypothesized to be due to the ability of the ncRNA to down-regulate [42] corepressor functions of Tup proteins [43, 44].

Another example of the above enhancement is work in embryonic stem cells of the transcription factor Ying Yang 1 (YY1) [41]. A number of pieces of evidence pointed to similar enhanced transcription in the presence of ncRNA. For example, artificially tethering RNA near YY1 binding sites, increased YY1 occupancy. These results suggest that ncRNA that is transcribed in proximity to YY1 acts to enhance further ncRNA transcription in this region.

Various models have been proposed [24] on how this enhancement could take place, including the trapping of transcription factors by the ncRNA, the recruitment of proteins that increase transcription factor binding, and the inhibition of proteins the repress transcription factor binding. These mechanisms are quite general; they imply that an increase in ncRNA concentration around some particular regulatory elements, should enhance further ncRNA transcription. This increased activation by ncRNA is a known general function of it [6].

In the case studied here, the above enhancement mechanism can potentially lead to the desired behavior shown in Fig. 8. Binding of additional ncRNA will increase the local ncRNA concentration which, as argued above, will lead to an increased rate of ncRNA transcription. This is illustrated in Fig. 10. RNA polymerase (labeled RNAP), transcribes ncRNA from DNA. This ncRNA enhances the binding of a transcription factor (TF). Binding of more ncRNA will further increase transcription due to the increased concentration of ncRNA near TF. Note that this mechanism is measuring bound ncRNA, rather than measuring unbound ncRNA. More specifically, it is measuring the probability of binding for an individual ncRNA molecule that is being transcribed. This gives a measure of precisely what we want, the ratio of bound ncRNA to its total concentration for species i, which equals $1 - \rho_i/C_i$. The fact that it correctly divides ρ_i by the total concentration, and is similar to known enhancement mechanisms, makes it a promising direction to consider.

7. Conclusions

Genetic networks are extremely complex and individual pathways have taken years of study to elucidate. It is quite apparent by now that ncRNA plays an important role and is not just "junk" as had been previously hypothesized [3, 4, 5]. The purpose of this work is not hypothesize yet another theoretical model that can be added to the list of potential mechanisms that biology may be using. Instead, it is to take a step back and look for new paradigms that can be used to understand genetic regulation.

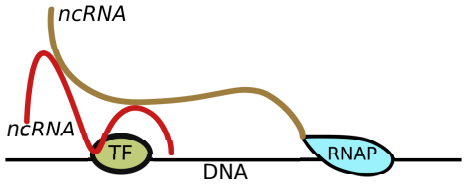


Figure 10: Possible mechanism for enhanced transcription due to bound RNA. RNA polymerase II (RNAP) transcribes DNA producing ncRNA. The presence of RNA enhances the binding of a transcription factor (TF), further enhancing the transcription of ncRNA.

Instead of thinking of interactions between individual elements as having a substantial and measurable effect on each other's behavior, the view taken here, is that there is a class of interactions, each of which is negligible, but collectively they are able to control behavior, perhaps even more effectively, than the sparse network models currently employed. Of course there are many examples where a single interaction has a large effect on gene expression, however here I examined if collective regulation could also play a substantial role.

Collective regulation is certainly possible mathematically, and has the same general architecture that is now used in machine learning systems [12]. What I have argued here is that this is also biologically and physically plausible. It is certainly the case that there are strong specific interactions that are involved in many regulatory pathways. However there is also a large amount of ncRNA that is highly associative, and is not very specific. The approach taken here is to accept the existence of thousands (or millions) of potential interactions between different RNA species and understand how these could evolve from junk, inserted by retroviruses, to become useful additions to the cell's genome.

The interactions between the RNA molecules in equilibrium yield a chemical equilibration between bound and unbound states. In reality there will be many higher order interactions and different internal states of molecules. These will surely affect the way computation takes place in these systems, but would not necessarily diminish their computational capabilities. Similarly, models of neurons that only consider two-body interactions, such as Boltzmann machines, leave out a lot of higher body effects that are present with real neurons. In this sense, the work here is only really a proof of principle. If collective regulation does exist, it would clearly be more complex than described here.

The chemical equilibration formula, Eq. 5, contains the seed of how this system is related to artificial neural network models, by making an analogy with Eq. 1. The equilibrium constants K_{ij} are analogous to the interactions J_{ij} between different "spins" in neural networks. It is the sum of all unbound RNA, weighted with equilibrium constants, that self consistently must give back the correct concentration of unbound RNA.

The constant degradation of RNA can be taken into account with a first order reaction rate equation, Eq. 6. But there also needs to be a mechanism for the replenishment of RNA. First a homeostatic mechanism needs to be included

to regulate the total concentration of all RNA. But the most difficult part of the analysis, that I investigated analytically and numerically, is that the creation rate of the *ith* species should only be a function of the ratio of that RNA's total concentration to the amount that is unbound, C_i/ρ_i. Such a function should look similar to what is shown in Fig. 8. I was able to show for a large class of interactions, that this function can be universal, in that it only depends C_i/ρ_i, and with no dependence on the particular species of RNA being created.

The physical binding and unbinding of ncRNA should happen according to estimates using empirical data discussed in Sec. 6.1. RNA species creation are also subject to a variety of regulatory mechanisms. The validity of the proposal outlined here then boils down to whether there exists RNA creation rates in the nucleus that depend on C_i/ρ_i according to Fig. 8. I argued in Sec. 6.2 that in fact, there is evidence for similar mechanisms already. This does not prove the existence of the kind of computational hardware discussed here, but argues that it is at least plausible. Further discussion of the issue of its evolutionary likelihood is unlikely to be fruitful and there appears to be many schools of thought on this issue [45].

These kinds of computational paradigms have several advantages, one being that they are far more robust than circuitry with few connections [12], and this would mean that one would expect ncRNA would have a much higher mutation rate than mRNA, yet be highly functional. If this kind of collective regulation does exist, it would imply a reexamination of how mutation rate can be used as a criterion for when ncRNA is under evolutionary constraint. It should also be noted that this feature of high mutation rate makes such massive parallelism unlikely in protein regulatory networks, which also have clear similarities with neural networks [46].

This mechanism also ties in with work to understand the molecular evolution of the genome, for example, transposable elements, which constitute approximately 44% of our genome, as evident from the Human Genome Browser [47, 48]. It would be of interest to consider the beneficial effects of the ncRNA discussed here to see their effects of population genetics simulations of them [49].

In this kind of collective mechanism, one expects that typically there will be weak influences between any two RNA molecules, and also very many such weak interactions. This makes it difficult to reconstruct the circuit diagram, in contrast to sparser networks where powerful methods exist [50]. In addition, molecules that are involved with this kind of collective regulation could have other functions, making it hard to identify specific interactions that support this picture. The understanding of ncRNA-ncRNA interactions is still in its infancy.

This work was supported by the Foundational Questions Institute <http://fqxi.org>.

[1] E. P. Consortium, et al., Identification and analysis of functional elements in 1% of the human genome by the encode pilot project, nature 447 (7146) (2007) 799.

[2] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willing-

ham, P. F. Stadler, J. Hertel, J. Hackermüller, I. L. Hofacker, et al., Rna maps reveal new rna classes and a possible function for pervasive transcription, Science 316 (5830) (2007) 1484–1488.

[3] T. R. Mercer, M. E. Dinger, J. S. Mattick, Long non-coding rnas: insights into functions, Nature Reviews Genetics 10 (3) (2009) 155.

[4] E. P. Consortium, et al., An integrated encyclopedia of dna elements in the human genome, Nature 489 (7414) (2012) 57.

[5] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, et al., Landscape of transcription in human cells, Nature 489 (7414) (2012) 101.

[6] J. T. Lee, Epigenetic regulation by long noncoding rnas, Science 338 (6113) (2012) 1435–1439.

[7] J. Deutsch, Collective regulation by non-coding rna, arXiv preprint arXiv:1409.1899.

[8] E. Mjolsness, D. H. Sharp, J. Reinitz, A connectionist model of development, Journal of theoretical Biology 152 (4) (1991) 429–453.

[9] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, et al., The nhgri gwas catalog, a curated resource of snp-trait associations, Nucleic acids research 42 (D1) (2013) D1001–D1006.

[10] P. M. Visscher, M. A. Brown, M. I. McCarthy, J. Yang, Five years of gwas discovery, The American Journal of Human Genetics 90 (1) (2012) 7–24.

[11] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, et al., Common snps explain a large proportion of the heritability for human height, Nature genetics 42 (7) (2010) 565.

[12] J. Hertz, A. Krogh, R. G. Palmer, Introduction to the theory of neural computation, Redwood City CA: Addison-Wesley.

[13] D. E. Rumelhart, J. L. McClelland, Parallel Distributed Processing, Explorations in the Microstructure of Cognition: Foundations Volume 1, MIT Press, Cambridge Mass., 1986.

[14] W. A. Little, Little, w. a. the existence of persistent states in the brain. math. biosci. 19, 101-120 19 (1974) 101–120.

[15] W. Little, G. L. Shaw, A statistical theory of short and long term memory, Behavioral Biology 14 (2) (1975) 115 – 133.

[16] W. Little, G. L. Shaw, Analytic study of the memory storage capacity of a neural network, Mathematical Biosciences 39 (3) (1978) 281 – 290.

[17] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proceedings of the national academy of sciences 79 (8) (1982) 2554–2558.

[18] D. J. Amit, H. Gutfreund, H. Sompolinsky, Spin-glass models of neural networks, Physical Review A 32 (2) (1985) 1007.

[19] D. J. Amit, H. Gutfreund, H. Sompolinsky, Information storage in neural networks with low levels of activity, Physical Review A 35 (5) (1987) 2293.

[20] J. Deutsch, Associative memory by collective regulation of non-coding rna, arXiv preprint arXiv:1608.05494.

[21] W. Poole, A. Ortiz-Munoz, A. Behera, N. S. Jones, T. E. Ouldridge, E. Winfree, M. Gopalkrishnan, Chemical boltzmann machines, in: International Conference on DNA-Based Computers, Springer, 2017, pp. 210–231.

[22] S. Spiegelman, I. Haruna, I. Holland, G. Beaudreau, D. Mills, The synthesis of a self-propagating and infectious nucleic acid with a purified enzyme, Proceedings of the National Academy of Sciences 54 (3) (1965) 919–927.

[23] P. Kapranov, F. Ozsolak, S. W. Kim, S. Foissac, D. Lipson, C. Hart, S. Roels, C. Borel, S. E. Antonarakis, A. P. Monaghan, et al., New class of gene-termini-associated human rnas suggests a novel rna copying mechanism, Nature 466 (7306) (2010) 642.

[24] N. Takemata, K. Ohta, Role of non-coding rna transcription around gene regulatory elements in transcription factor recruitment, RNA biology 14 (1) (2017) 1–5.

[25] G. E. Hinton, T. J. Sejnowski, Analyzing cooperative computation, in: Proceedings of the Fifth Annual Conference of the Cognitive Science Society, Rochester NY, 1983.

[26] G. E. Hinton, T. J. Sejnowski, Optimal perceptual inference, in: Proceedings of the *IEEE* Conference on Computer Vision and Pattern Recognition, 1983, p. 448453.

[27] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by simulated annealing, science 220 (4598) (1983) 671–680.

[28] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, J. L. Rinn, Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses, Genes & development 25 (18) (2011) 1915–1927.

[29] N. E. Buchler, U. Gerland, T. Hwa, On schemes of combinatorial transcription logic, Proceedings of the National Academy of Sciences 100 (9) (2003) 5136–5141.

[30] F. Reif, Fundamentals of statistical and thermal physics, Waveland Press, 2009.

[31] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, et al., The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression, Genome research 22 (9) (2012) 1775–1789.

[32] R. Milo, R. Phillips, Cell biology by the numbers, Garland Science, 2015.

[33] J. C. Politz, T. Pederson, Movement of mrna from transcription site to nuclear pores, Journal of structural biology 129 (2-3) (2000) 252–257.

[34] E. A. Lesnik, S. M. Freier, Relative thermodynamic stability of dna, rna, and dna: Rna hybrid duplexes: relationship with base composition and structure, Biochemistry 34 (34) (1995) 10807–10815.

[35] M. Turner, A. Galloway, E. Vigorito, Noncoding rna and its associated proteins as regulatory elements of the immune system, Nature immunology 15 (6) (2014) 484.

[36] E. Sharma, T. Sterne-Weiler, D. OHanlon, B. J. Blencowe, Global mapping of human rna-rna interactions, Molecular cell 62 (4) (2016) 618–626.

[37] J. M. Engreitz, K. Sirokman, P. McDonel, A. A. Shishkin, C. Surka, P. Russell, S. R. Grossman, A. Y. Chow, M. Guttman, E. S. Lander, Rna-rna interactions enable specific targeting of noncoding rnas to nascent pre-mrnas and chromatin sites, Cell 159 (1) (2014) 188–199.

[38] S. Akira, Toll-like receptors and innate immunity, Advances in immunology 78 (2001) 1–56.

[39] L. Alexopoulou, A. C. Holt, R. Medzhitov, R. A. Flavell, Recognition of double-stranded rna and activation of nf-κb by toll-like receptor 3, Nature 413 (6857) (2001) 732.

[40] N. Agrawal, P. Dasaradhi, A. Mohmmed, P. Malhotra, R. K. Bhatnagar, S. K. Mukherjee, Rna interference: biology, mechanism, and applications, Microbiology and molecular biology reviews 67 (4) (2003) 657–685.

[41] A. A. Sigova, B. J. Abraham, X. Ji, B. Molinie, N. M. Hannett, Y. E. Guo, M. Jangi, C. C. Giallourakis, P. A. Sharp, R. A. Young, Transcription factor trapping by rna in gene regulatory elements, Science 350 (6263) (2015) 978–981.

[42] N. Takemata, A. Oda, T. Yamada, J. Galipon, T. Miyoshi, Y. Suzuki, S. Sugano, C. S. Hoffman, K. Hirota, K. Ohta, Local potentiation of stress-responsive genes by upstream noncoding transcription, Nucleic acids research 44 (11) (2016) 5174–5189.

[43] Y. Mukai, E. Matsuo, S. Y. Roth, S. Harashima, Conservation of histone binding and transcriptional repressor functions in a schizosaccharomyces pombe tup1p homolog, Molecular and cellular biology 19 (12) (1999) 8461–8468.

[44] R. Asada, N. Takemata, C. S. Hoffman, K. Ohta, K. Hirota, Antagonistic controls of chromatin and mrna start site selection by tup family corepressors and the ccaat-binding factor, Molecular and cellular biology 35 (5) (2015) 847–855.

[45] J. J. Welch, Whats wrong with evolutionary biology?, Biology & philosophy 32 (2) (2017) 263–279.

[46] D. Bray, Protein molecules as computational elements in living cells, Nature 376 (6538) (1995) 307.

[47] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at ucsc, Genome research 12 (6) (2002) 996–1006.

[48] R. E. Mills, E. A. Bennett, R. C. Iskow, S. E. Devine, Which transposable elements are active in the human genome?, Trends in genetics 23 (4) (2007) 183–191.

[49] T. Kijima, H. Innan, Population genetics and molecular evolution of dna sequences in transposable elements. i. a simulation framework, Genetics 195 (3) (2013) 957–967.

[50] A. Mochizuki, B. Fiedler, G. Kurosawa, D. Saito, Dynamics and control at feedback vertex sets. ii: A faithful monitor to determine the diversity of molecular activities in regulatory networks, Journal of theoretical biology 335 (2013) 130–146.