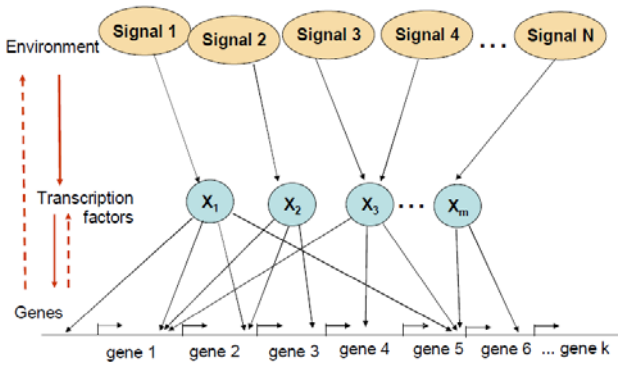
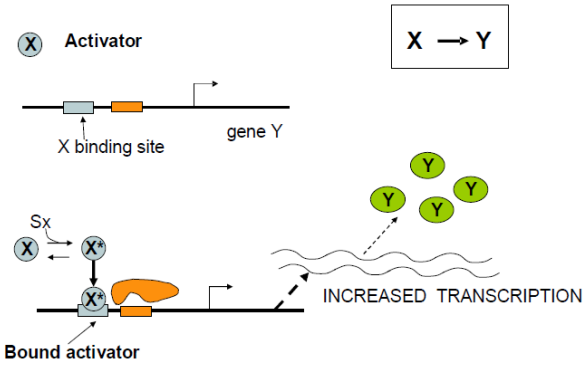
Chapter 3 Basic Concepts of Transcriptional Regulatory Networks and Relevant Motifs

3.1Basic Concepts of Transcriptional Regulatory Networks (TRN)

Cells are very complex systems, which have to communicate with each other and adapt to a changing environment for survival. It continuously monitors its environment and calculates the amount at which each type of protein is needed. This information processing function is carried out by transcription networks. To represent these environmental states, the cell uses special proteins called transcription factors. Transcription is the process of creating a complementary RNA copy of a sequence of DNA. RNA polymerase reads a DNA sequence and produces an antiparallel mRNA strand. The rate of transcription and the number of mRNAs produced is controlled by the promoter (a regulatory region of DNA that precedes the gene). The input signals $\{S_i\}$ carry information from the environment.



A transcription-factor protein can be an activator ($X \rightarrow Y$, positive control) that increase the rate of mRNA transcription when it binds the promoter. The activator transits rapidly between active and inactive forms. In its active form, it has a high affinity to a specific site (or sites) on the promoter. An input signal S_X can increase the probability that X is in its active form X*. Thus, X* binds the promoter of gene Y to increase transcription and production of protein Y. The time scales are typically sub second for transitions between $X \rightarrow X^*$, seconds for binding/unbinding of X to the promoter, minutes for transcription and translation of the protein product, and tens of minutes for the accumulation of the protein. A signal can cause X to rapidly shift to its active state X^*. A transcription-factor protein can also be a repressor ($X \dashv Y$, negative control) that decreases mRNA transcription when it binds the promoter. The signal S_X increases the probability that X is in its active form X*. X* binds a specific site in the promoter of gene Y to decrease transcription and production of protein Y.



The set of interaction between genes and the transcription factors forms a transcription network, which describes all regulatory transcription interactions in a cell. In the network, the nodes are genes and edges represent transcriptional regulation of one gene by the protein product of another gene. A directed edge $X \to Y$ means that the product of gene *X* is a transcription factor protein that binds the promoter of gene *Y* to control the rate at which gene *Y* is transcribed. Typically, positive control is more common, occupies 60 – 80% of interactions in E. Coli & yeast.

3.2 The Input Function

3.2.1 The Hill Function

The strength of the interaction associated with an edge of the network is described by an input function. Consider $X \to Y$ with X^* denoting the concentration of the active form of *X*, the production rate of $Y = f(X^*)$, where *f* is the input function.

An appropriate form for activators is the Hill function defined as

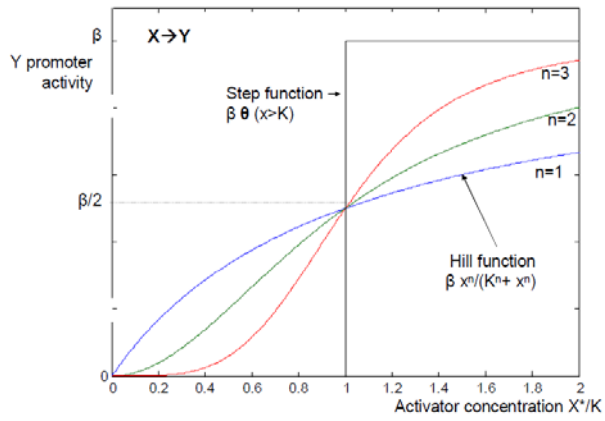
$$Y = f(X^*) = \frac{\beta X^{*n}}{K^n + X^{*n}}.$$

Here *K* is activation coefficient, denoting the concentration of active *X* that is needed to significantly activate expression of *Y*; thus half-maximal expression is reached when *X** = *K*. Maximal expression level β of *Y* can be reached when *X** >> *K*. Hill coefficient *n*: governs the steepness of the input function; typically *n* = 1, ..., 4. For repressors a proper form of Hill input function can be found to be:

$$Y = f(X^*) = \frac{\beta}{1 + (X^*/K)^n}.$$

Maximal expression level β of *Y* is reached when *X**=0.

The parameters for the input function can readily be tuned during evolution of an organism. Lab experiments have shown that when placed in new environment, bacteria can tune these parameters within hundreds of generations to reach optimal expression levels. Many genes have a nonzero minimal expression level, called the basal expression level, which can be modeled by adding term β_0 to the input function.

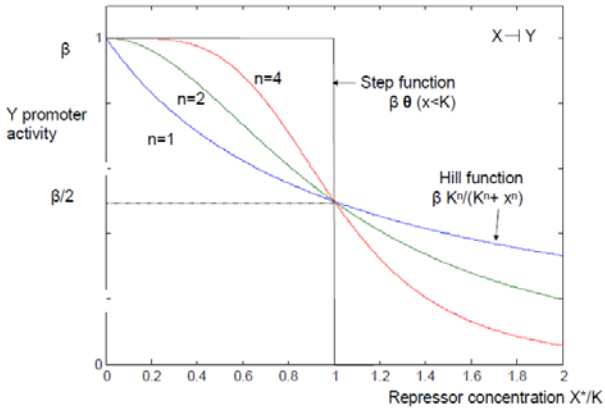


3.2.2 The Logic Input Function

For mathematical clarity, we often use simpler input functions. The essence of input function is transition between low and high values, with a characteristic threshold *K*. Using logic approximation the gene is either OFF with $Y = f(X^*) = 0$, or maximally ON $Y = f(X^*) = \beta$. Use the step-function

$$\Theta(z) = \begin{cases} 0, if\ \ z = false \\ 1, if\ \ z = true \end{cases},$$

the logic approximation for activator leads to $Y = f(X^*) = \beta\Theta(X^* > K)$, and $Y = f(X^*) = \beta\Theta(X^* < K)$ for repressor.



3.2.3 Multi-dimensional Input Function

The multi-dimensional input functions for genes with several inputs can be constructed for the AND function as

$$Y = f(X^*, Y^*) = \beta\Theta(X^* > K_X)\Theta(Y^* > K_Y) \sim X^*\ AND\ Y^*;$$

OR function as

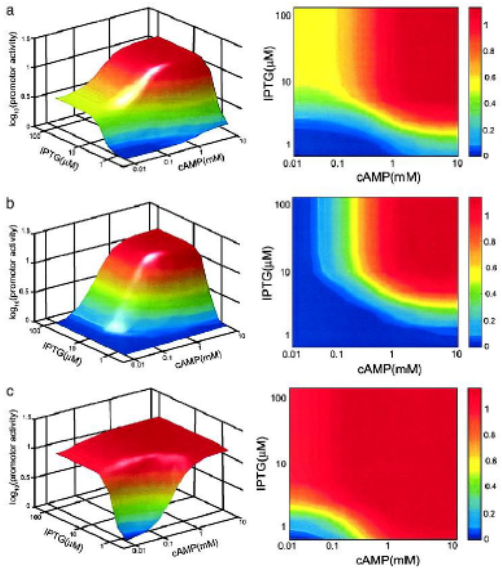
$$Y = f(X^*, Y^*) = \beta\Theta(X^* > K_X\ or\ Y^* > K_Y) \sim X^*\ OR\ Y^*;$$

SUM as

$$Y = f(X^*, Y^*) = \beta_X X^* + \beta_Y Y^*.$$

It appears that the precise form of the input function of each gene is under selection pressure during

evolution.



Two-dimensional input functions. (a) Input function measured in the lac-promoter of E. coli, as a function of two inducers cAMP and IPTG. (b) an AND-gate like input function, which shows high promoter activity only if both inputs are present. (c) an OR-gate like input function that shows high promoter activity if either input is present. Source: Setty, Y. *et al*. (2003) Proc. Natl. Acad. Sci. USA 100, 7702-7707.

3.3 Dynamics of Gene Regulation Networks

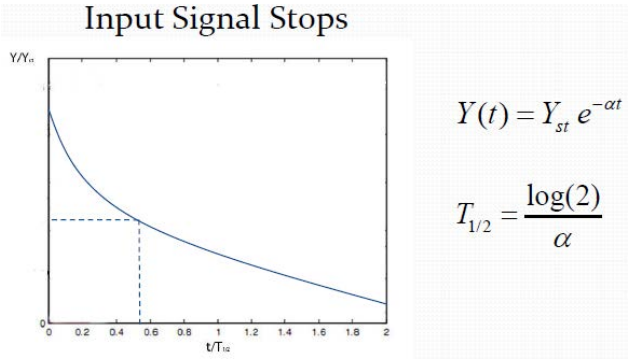
3.3.1 Response Time of a Simple Gene Regulation

Consider a simple gene regulation $X \rightarrow Y$ (that is Y is regulated only by X). A cell produces Y at a constant rate β (in units: concentration/time).The concentration of Y is balanced by degradation of Y by specialized proteins in the cell; and dilution (reduction in concentration of Y due to the increase of cell volume during cell growth). The total decay rate $\alpha = \alpha_{\deg} + \alpha_{dil}$ is the sum of the degradation rate $\alpha_{\deg}$ (in units: 1/time) and the dilution rate α_{dil} (in units: 1/time), leads to the change in Y concentration as:

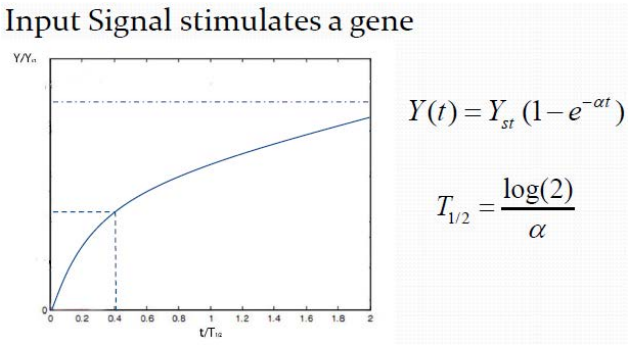
$$\frac{dY}{dt} = \beta - \alpha Y .$$

(i) The equation predicts a steady state with $\frac{dY}{dt} = 0 \rightarrow Y_{st} = \beta/\alpha$.

(ii) When $Y = Y_{\text{st}}$ and then we take away the input signal S_{X} (hence β=0), Y will decay exponentially in t as $Y(t) = Y_{st}\, e^{-\alpha t}$. The response time $T_{1/2}$, which is defined as the time needed to reach halfway between the initial and final states in a dynamic process, is given as $T_{1/2} = \ln 2/\alpha$, which depends only on α.



(iii) What happens if a gene, which is unstimulated (*i.e.*, $Y = 0$), becomes suddenly stimulated by a strong signal S_X? The differential equation with a proper initial condition yields $Y(t) = Y_{st}(1 - e^{-\alpha t})$, which predicts the level of Y will start to accumulate towards the steady state concentration Y_{st}. The response time to reach $Y_{st}/2$ is therefore $T_{1/2} = \ln 2/\alpha$, which is the same for both decay and accumulation.



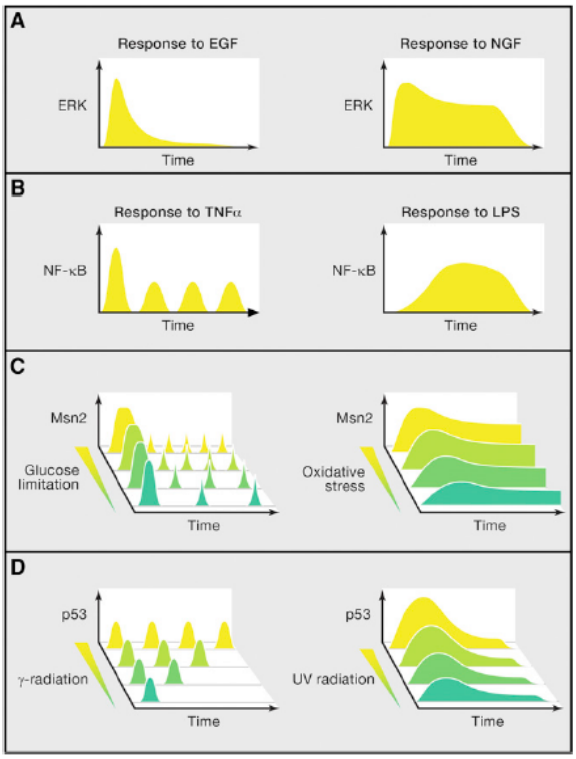
3.3.2 Dynamical Patterns of Cellular Signaling Networks

A) One of the intriguing responses of cellular signaling networks is that different upstream signals can lead to different dynamical patterns of the same network. An example of this behavior is the extracellular signal-regulated kinase (ERK) pathway. It was found that nerve growth factor (NGF) can lead to differentiation, whereas epidermal growth factor (EGF) results in cell proliferation. Both stimuli activate ERK but with distinct dynamical patterns. Specifically, EGF triggers a transient response, whereas NGF induces sustained ERK activation.

B) Upstream signals can also be encapsulated in their dynamics. One example is the inflammatory pathway, which shows different inflammatory stimuli induce distinct temporal profiles of the transcription factor NF-κB. Under resting conditions, NF-κB is continuously shuttled between nuclear and cytosolic compartments. Activation of NF-κB by tumor necrosis factor-α (TNFα) results in prolonged occupation in the nucleus and transcription of its negative regulator IκBα. This negative feedback loop generates oscillations of active NF-κB. In contrast, bacterial lipopolysaccharide (LPS) leads to slower accumulation and a single prolonged wave of NF-κB activity.

C) Both the identity and strength of the stimulus can alter the dynamics of a signaling protein. One example is the yeast transcription factor Msn2, which responds to stress by translocation to the nucleus. In response to glucose limitation or high osmolarity, nuclear Msn2 shows a transient

increase with a dose-dependent duration and fixed amplitude. Following the initial pulse, glucose limitation and osmotic stress lead to a series of Msn2 bursts. The frequency of these pulses depends on the signal intensity of glucose limitation but is not affected by the intensity of the osmotic stress. In contrast, oxidative stress leads to prolonged nuclear Msn2 accumulation with amplitude that increases with higher concentration of H_2O_2.



D) The tumor suppressor p53 also shows both stimulus- and dose-dependent dynamics. Double-strand breaks (DSBs) caused by γ-radiation trigger a series of p53 pulses with fixed amplitude and duration. Higher doses of radiation increase the number of pulses without affecting their amplitude or duration. In contrast, UV triggers a single p53 pulse with a dose-dependent amplitude and duration.

The identification of network motifs in transcription networks and the study of their dynamics in various systems have revealed a strong relationship between motif structure, dynamics, and specific function. Further understanding how dynamics are regulated at the molecular level certainly will become the major subject of the incoming decade.

3.4 Network Motifs

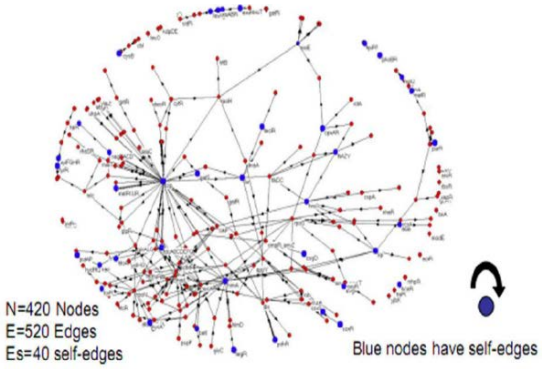
Network motifs in a real network can be discovered by comparing the network to an ensemble of randomized networks pattern and to find out the connection patterns that occur significantly more often than in the randomized networks. For a meaningful comparison, the randomized networks should share the basic features of the real network –same number of nodes and edges.

Using the direct transcriptional interactions in *Escherichia coli* as an example, much of the network is composed of repeated appearances of three highly significant motifs. Each network motif has a specific function in determining gene expression, such as generating temporal expression programs, governing the responses to fluctuating external signals, or providing a memory effect.

3.4.1 The First Network Motif: Autoregulation

The first motif, termed 'autoregulation' occurs when a transcription factor *X* provides a regulation on the transcription of its own gene *X* with a schematic shown as 

. The regulation effect can be either negative (Negative AutoRegulation NAR) or positive (Positive AutoRegulation PAR).



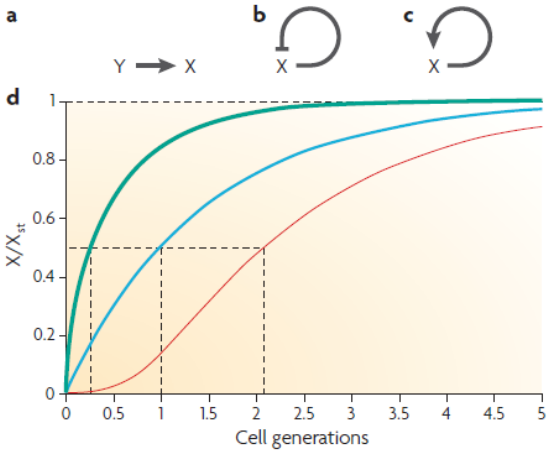
As shown above, the *E. Coli* transcription network has 424 nodes, 519 edges and 40 self edges. Among the 40 self edges, 34 repress their own transcription. In a random network with the same number of nodes *N* and edges *E*, a self edge can occur in the network with a probability of $P_{se} = 1/N$. Since *E* edges are placed at random to form the random network, the probability of having *k* self-edges is approximately binomial $P(k) = \binom{E}{k} P_{se}{}^{k} (1 - P_{se})^{E-k}$. Thus the expected number of self-edges in the random network becomes

$$< N_{se} >_{rand} = \sum_{k=1}^{E} kP(k) \sim EP_{se} \sim \frac{E}{N} = \frac{520}{420} \sim 1.2 .$$

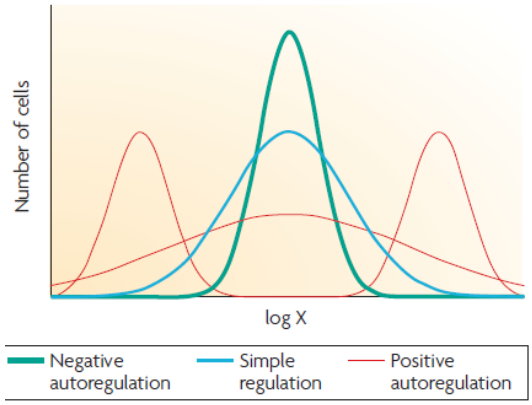
The standard deviation of the number of self edges is $\sigma_{rand} = \sqrt{E/N} \sim 1.1$. In contrast, the *E. Coli* transcription network has 40 self-edges, which exceeds the random networks by many standard deviations. We can describe the significant difference in the number of self-edges as $Z = (< N >_{real} - < N >_{rand})/\sigma_{rand} = (40 - 1.2)/1.1 \sim 32$, which means they occur far more in the real network than at random . Thus self edges are a network motif---AR.

NAR can speed up the response time of gene circuits. This occurs when NAR uses a strong promoter to obtain a rapid initial rise in the concentration of protein X. As shown earlier , the dynamics of *X* with NAR can be described by its production rate and degradation/dilution rate as $dX/dt = f(X) - \alpha X = \beta/[1 + (X/K)^n] - \alpha X$ .To solve this in the most intuitive way, we will use the logic approximation of *f*(*X*) as $f(X) = \beta\Theta(X < K)$. To study the response time, consider the case where *X* is initially absent, and its production starts at *t*=0. At early time, while *X* concentration is low, the promoter is unrepressed and production is at full rate $dX/dt = \beta - \alpha X \ as \ X < K$. This yields an initial rising *X* as $X(t) = \beta t \ as \ X < K \ \ and \ X < X_{st} = \beta/\alpha$. On the other hand, production stops when *X* levels reach the self-repression threshold, *X*=*K*. Thus, *X* locks into a steady-state level X_{st},

which is equal to the repression coefficient of its own promoter $X_{st} = K$. Thus $X(T_{1/2}) = \beta T_{1/2} = K/2$ gives $T_{1/2}^{NAR} = K/(2\beta)$. The response time is shorter in the case of NAR (green curve) than that with simple regulation (blue line).



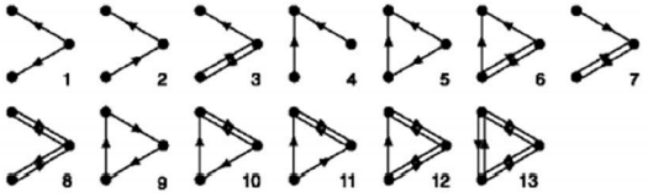
In a cell, the production rate of a protein can fluctuate by tens of percents, which can cause a significant cell–cell variation. NAR can reduce the cell–cell variation in protein levels: high concentrations of *X* reduce its own rate of production, whereas low concentrations cause an increased production rate. The result is a narrower distribution (green curve) of protein levels than would be expected in simply regulated genes (blue line).



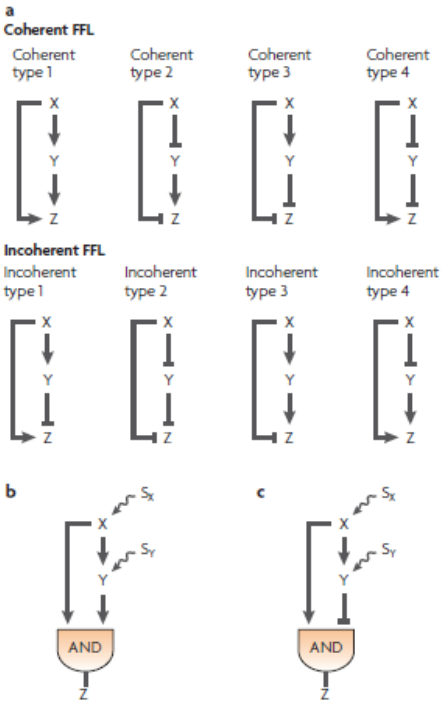
The effects of positive autoregulation are opposite to those of NAR: response times are slowed and variation is usually enhanced. PAR slows the response time because at early stages, when levels of *X* are low, production is slow. Production picks up only when *X* concentration approaches the activation threshold for its own promoter. Thus, the desired steady state is reached in an *S*-shaped curve (red line). PAR tends to increase cell–cell variability. If PAR is weak, the cell–cell distribution of *X* concentration is expected to be broader than in the case of a simply regulated gene. Strong PAR can lead to bimodal distributions (red line), whereby the concentration of *X* is low in some cells but high in others.

3.4.2 The Second Network Motif: Feedforward Loop

There are 13 possible three node patterns. Using the random network model we can calculate the number of appearances of each sub graph, and compare to the transcription *E.Coli* network.



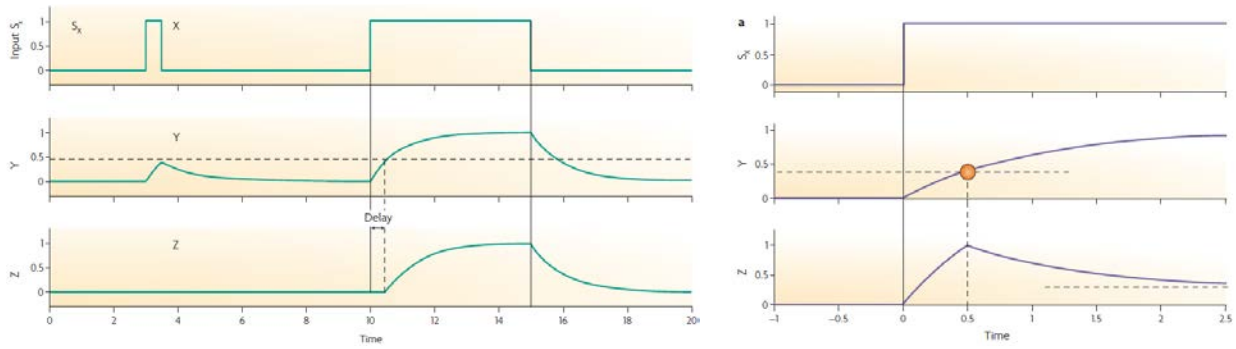
The result of the analysis is that there are 42 sub graph 5 in the transcription E.Coli network, whereas in the corresponding random network there are only ~1.2. This second motif, termed 'feedforward loop' (FFL) defined by a transcription factor *X* (the 'general transcriptional factor') that regulates a second transcription factor *Y* (the 'specific transcription factor'), such that both *X* and *Y* jointly regulate an operon Z (the 'effector operon(s)').



To understand the function of the FFLs, we need to understand how *X* and *Y* are integrated to regulate the *Z* promoter. Two common 'input functions' are an 'AND gate', in which both *X* and *Y* are needed to activate *Z*, and an 'OR gate', in which binding of either regulator is sufficient. Other input functions are possible, such as the additive input function. In the transcriptional networks of *E. coli* and *yeast*, two of the eight FFL types occur much more frequently than the other six types. These common types are the coherent type-1 FFL (C1-FFL, , both *X* and *Y* are transcriptional activators) and the incoherent type-1 FFL (I1-FFL, the two arms of the FFL act in opposition).

The C1-FFL is a 'sign-sensitive delay' element and a persistence detector. At time t=0, S_X triggers the activation of *X*. *X* rapidly transits to its active form *X** and binds to the promoters of genes *Y* and *Z*. Following an ON step of S_x , *Y** begins to be produced exponentially and then converge to a steady state level $Y^*(t) = Y_{st}(1-e^{-\alpha_y t}) = \frac{\beta_y}{\alpha_y}(1-e^{-\alpha_y t})$. *Z* production is governed by an AND input function of *X** and *Y**, yielding a delay after stimulation, but no delay when stimulation stops.

This dynamic behavior is called sign-sensitive delay; that is, delay depends on the sign of the S_x step. An ON step causes a delay in *Z* expression, but an OFF step causes no delay. The delay generated by the FFL can be useful to filter out brief spurious pulses of signal. A signal that appears only briefly does not allow *Y* to accumulate and cross its threshold, and thus does not induce a *Z* response. Only persistent signals lead to *Z* expression.



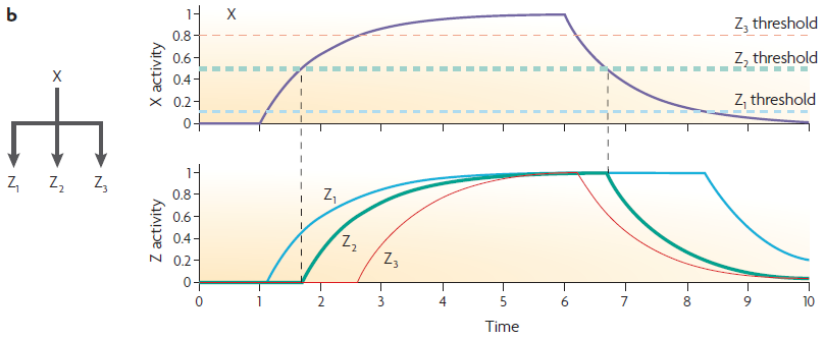
In the I1-FFL, the two arms of the FFL act in opposition: *X* activates *Z*, but also represses *Z* by activating the repressor *Y*. As a result, when a signal causes *X* to assume its active conformation, *Z* is rapidly produced. However, after some time, *Y* levels accumulate to reach the repression threshold for the *Z* promoter. As a result, *Z* production decreases and its concentration drops, resulting in pulselike dynamics. Thus, the I1-FFL can serve as a pulse generator and response accelerator.

When the *Z* promoter has OR logic, the FFL has the opposite effect to the AND case: the C1-FFL shows no delay after stimulation, but does show a delay when stimulation stops.

3.4.3 The Third Network Motif: Single-Input Module

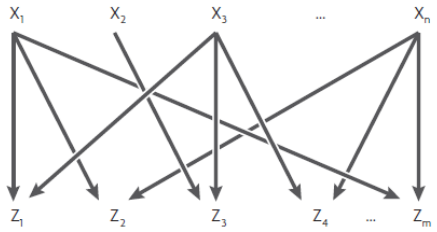
The third motif, termed single-input module (SIM), is defined by a set of operons that are controlled by a single transcription factor. All of the operons are under control of the same sign (all positive or all negative) and have no additional transcriptional regulation. The transcription factors controlling SIM motifs are usually autoregulatory (70%, mostly autorepression), in contrast to only 50% of the transcription factors in the complete data set.

The main function of this motif is to allow coordinated expression of a group of genes with shared function. It can generate a temporal expression program with a defined order of activation of each of the target promoters. *X* often has different activation thresholds for each gene, owing to variations in the sequence and context of its binding site in each promoter. So, when *X* activity rises gradually with time, it crosses these thresholds in a defined order, first the lowest threshold, then the next lowest threshold, an so on, resulting in a temporal order of expression



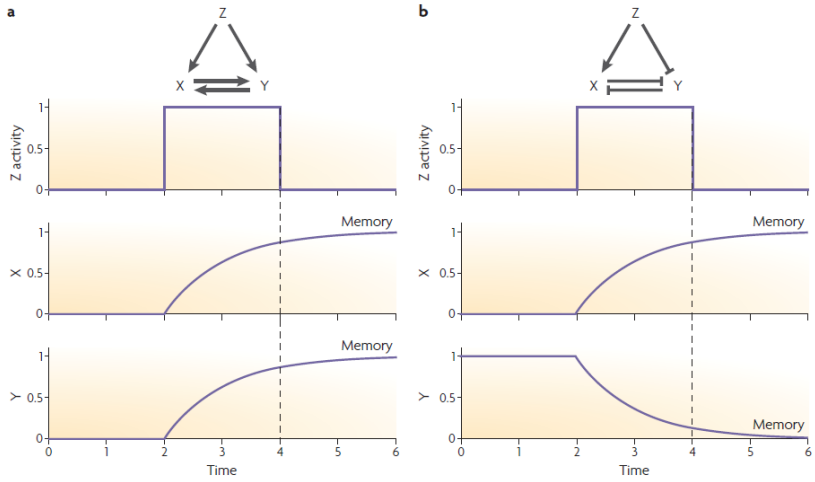
3.4.4 The Fourth Network Motif: Dense Overlapping Regulon

The fourth motif is referred to as dense overlapping regulons (DORs) or multi-input motifs (MIMs). *E. coli* has several DORs with hundreds of output genes, each responsible for a broad biological function, such as carbon utilization, anaerobic growth, stress response, and so on. The DOR can be thought of as a gate-array, carrying out a computation by which multiple inputs are translated into multiple outputs.



3.4.5 The Memory Module

Developmental transcription networks often use positive-feedback loops that are made up of two transcription factors that regulate each other. There are two kinds of positive feedback loops, (a) a double-positive loop and (b) a double-negative loop.



In the double-positive loop, two activators activate each other to exhibit two steady states: either both X and Y are OFF, or both are ON. The double-negative loop, in which two repressors repress each other, which generates different steady states: either X is ON and Y is OFF, or the opposite. In both cases, a transient signal can cause the loop to lock irreversibly into a steady state. In this sense, this network motif can provide memory of an input signal, even after the input signal is gone.