CHAPTER 1: INTRODUCTION

1.1 Biological networks in systems biology

Systems biology is a multidisciplinary discipline that aims to study biological system at the network level (Kitano, 2001). Currently, several types of biological networks are actively studied by researchers. These include transcriptional, metabolic, developmental, and signal transduction and neurophysiological networks (Duarte *et al.*, 2007; Fusco *et al.*, 2010; Le Novère, 2007; Pop *et al.*, 2010). Biological networks in both prokaryotes and eukaryotes are intensively studied. Many are carefully curated and deposited in publicly available databases such as KEGG (Kanehisa *et al.*, 2011), EcoCyc (Keseler *et al.*, 2005), RegulonDB (Gama-Castro *et al.*, 2011) and YMDB (Jewison *et al.*, 2012). These biological networks are often complex and difficult to use for downstream analyses such as identification of genes involved in lipid biosynthetic pathways (Hewald *et al.*, 2006). Despite the apparent complexity, it has been found that complex networks may nevertheless be made up of subnetworks of the same type (network motifs), thus opening up the possibility of predicting the behaviour of a complex network by analysing the behaviour of subnetworks (Alon, 2006).

1.2 Transcription network model

In a transcription network, there are typically two elements involved: transcription factors and genes (Alon, 2006). Figure 1.1 shows a simple transcriptional regulation of gene *Y* by a transcription factor *X*, which acts as an activator to initiate transcription of gene *Y*. An activation signal S_x activates *X*, causing it to bind the *X* promoter site of the target gene. Together with the binding of the RNA polymerase and other essential molecules to RNA polymerase binding site, the transcription of gene *Y* is triggered. Subsequently, the transcribed mRNA molecules are translated into Protein *Y* (Alon, 2006). Activation of *araBAD* by AraC in the arabinose system in *E. coli* is an example of activation of gene transcription by transcription factor in its active form. AraC is *ara* operon activator which triggers the transcription of *araBAD* when it is activated by its activation signal, L-arabinose (Lee *et al.*, 1973; Mangan *et al.*, 2003).

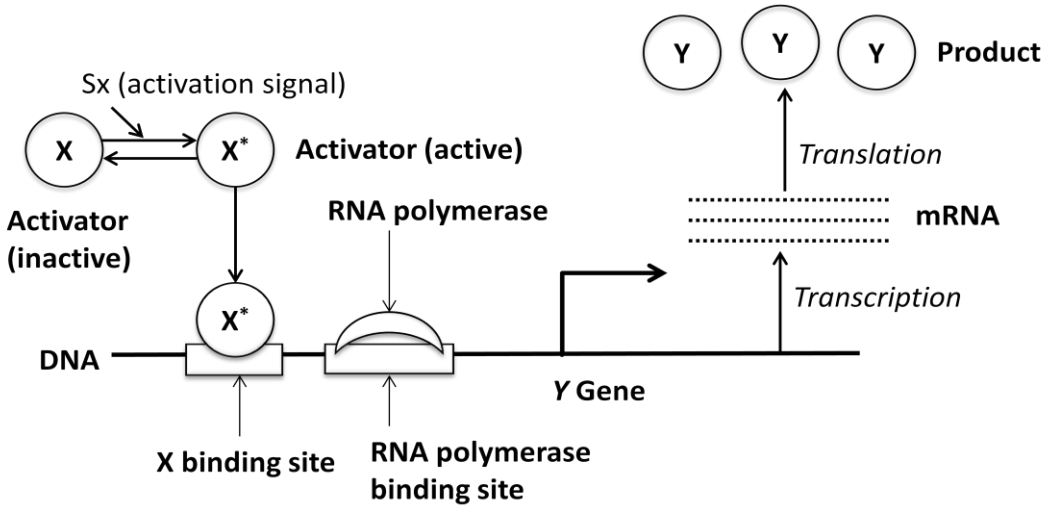

Figure 1.1: Transcriptional activation of gene *Y* by transcription factor *X*. (Alon, 2006)

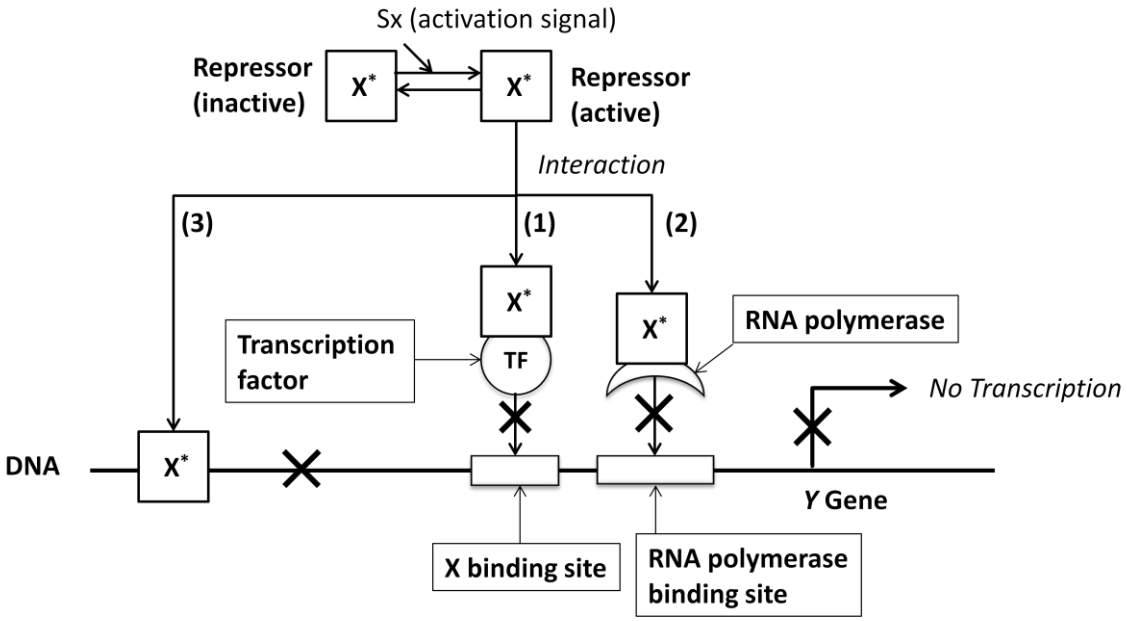

Figure 1.2: Three mechanisms of transcriptional repression of gene *Y* by transcriptional repressor *X*: (1) Re-organising the chromatin by recruiting chromatin-remodelling proteins; (2) interfering the binding of transcription factors to promoter sequences; and (3) modifying the structure of RNA polymerase (Gaston & Jayarama, 2003).

In certain cases, proteins in activated by its activation signal may act as transcriptional repressors which inhibit the transcription of the genes (Figure 1.2; Gaston & Jayarama, 2003). There are types of transcriptional repression: general/global repression and gene-specific repression. General transcriptional repressors inhibit transcription by modifying the structure of RNA polymerase. Specific transcriptional repressors inhibit transcription by interfering the binding of transcription factors to promoter sequences or re-organising the chromatin by recruiting chromatin-remodelling proteins. Repression of *galETK* by GalS in the galactose system in *E. coli* is an example of inhibition of gene transcription by a transcriptional repressor in its active form. GalS is a Gal isorepressor which inhibits transcription of *galETK* when it is activated by its activation signal, galactose or D-fucose (Csiszovszki *et al.*, 2011; Mangan *et al.*, 2006).

1.3 Network motifs as Individual building blocks of complex biological network complex

In order to analyse the biological functions of complex networks, a reductionist approach that break down the network into basic individual building block has yielded interesting results that can be experimentally validated (Alon, 2006). With the theoretical network model hypothesised by Erdős and Renyi in 1960, certain building block patterns of naturally biological network circuit appear to exist more often than intentional ensemble of randomised network (Erdős & Rényi, 1960). These repeating patterns of network are termed network motifs. Network motifs are simplified and useful model to understand the evolution of the biological networks as well as their significance in biological functions for the development and survival of organisms (Schramm *et al.*, 2010).

1.4 Types of transcription network motifs

Several types of transcription network motifs have been identified: simple regulation, single-input modules (SIM), dense overlapping regulons (DOR), and feed-forward loop (FFL) (Alon, 2007; Alon, 2006; Shen-Orr *et al.*, 2002). Simple regulation is further subdivided into positive autoregulation (PAR) and negative autoregulation (NAR) (Figure 1.3). In PAR, the gene product increases their own production by activating its own promoter, thus resulting in higher concentration of gene products than simple regulation. In contrast to PAR, the gene product in NAR represses the activity of its own promoter, thus maintaining the product concentration at an optimal physiological level.

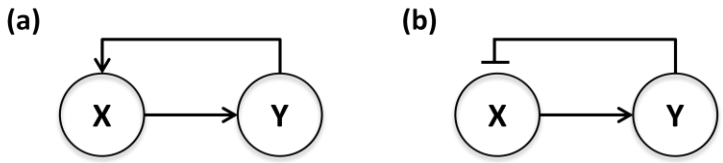

Figure 1.3: Types of simple regulation of gene *Y* by transcription factor *X*. (a) Positive autoregulation (PAR), and (b) negative autoregulation (NAR). (Alon, 2006)

Single-input modules (SIM) involve the regulation of multiple genes by a single type of transcription factor either by activation or repression (Figure 1.4). It usually functions to control a group of genes or operons in the case of prokaryotes with shared functions. The concentration of transcription factor itself may regulate its own promoter activity. Each regulation between the transcription factor and the target genes has different activation or repression threshold. The difference in activation or repression threshold results in different activation or repression time point based on the parameters of the activation or repression threshold.

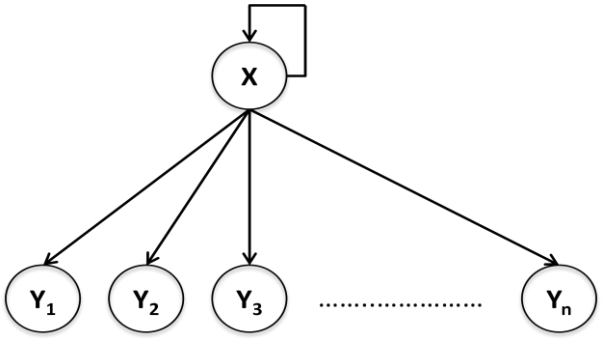

Figure 1.4: Single-input modules network motifs involved in regulation of multiple target genes Y_1, Y_2, Y_3,…Y_n by transcription factor *X* (Alon, 2006).

Dense overlapping regulons (DOR), also known as multi-input motifs (MIM), is a set of combinatorial interaction between multiple transcription factors and multiple genes (Figure 1.5). DOR network motifs, which are physiologically important, are found as transcription networks in *Escherichia coli* and *Saccharomyces cerevisiae*. The input and output functions are relatively complicated and unknown in comparison to other types of network motifs.

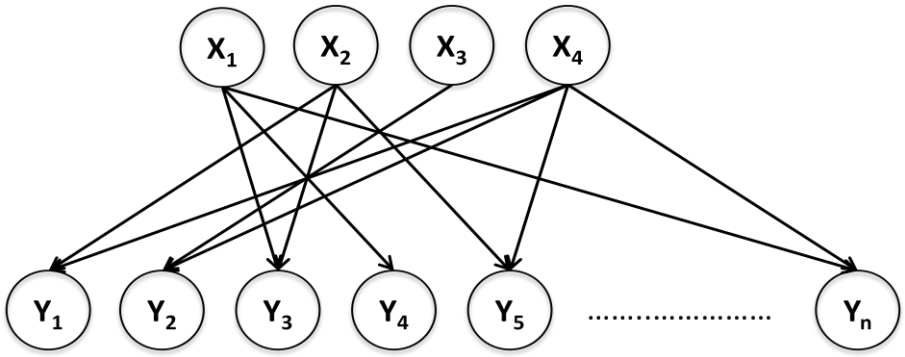

Figure 1.5: Single-input modules network motifs involved in regulation of multiple target genes Y_1, Y_2, Y_3,…Y_n by a set of transcription factors X_1, X_2, X_3, and X_4 (Alon, 2006).

1.5 Feed-forward loop

A feed-forward loop (FFL) is a simple, closed graph with three nodes connected by three unidirectional edges (Figure 1.6). The nodes represent transcriptional factors or genes, and the edges represent interaction type (activation or repression). Biological networks have been found to contain FFL as a network motif (Alon, 2006).

The cascade of reactions in an FFL requires the activation of the transcription factors by their respective activation signals. For example, *X* is activated by its activation signal S_x; it then directly regulates the transcription of both gene *Y* and *Z*. The product of gene *Y*, regulated by transcription factor *X*, and upon activation by its activation signal S_y, also acts as a transcription factor to regulate the transcription of gene *Z* simultaneously with transcription factor *X*.

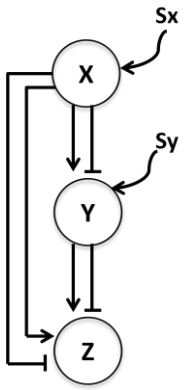

Figure 1.6: Feed-forward loop with transcriptional regulation by activation and/or repression. *X*, *Y* and *Z* are transcriptional factors. *X* activated by signal S_x regulates both *Y* and *Z* while *Y* activated by signal S_y regulates only *Z* (Alon, 2006).

There are totally eight types of FFL based on the regulatory patterns. Eight types of FFL can be categorised into two groups: coherent and incoherent FFL. In coherent FFL, the indirect path has the same regulatory type as the direct path (*X* to *Z*), while in incoherent FFL, the indirect path (*X* to *Y* to *Z*) has the opposite regulatory type as the direct path. These results in four types of FFL belonging to each of the coherent and incoherent FFL: type 1 to type 4 coherent FFL (C1-FFL to C4-FFL), and type 1 to type 4 incoherent FFL (I1-FFL to I4-FFL) (Figure 1.7).

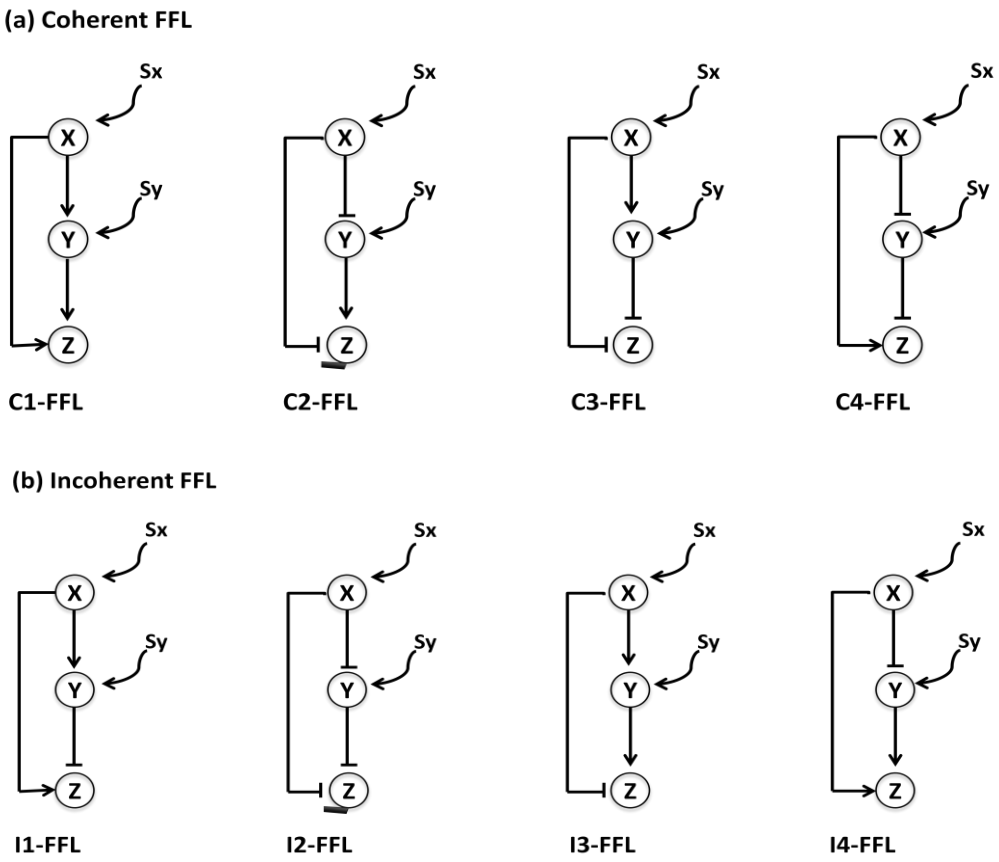

Figure 1.7: Eight types of FFL: (a) coherent FFL; (b) incoherent FFL (Alon, 2006).

The transcription of gene *Y* is regulated by the transcription factor *X* and *Y* following Boolean input logics: AND or OR gate (Figure 1.8). C1-FFL is used as an example to discuss the difference of the regulatory pattern of FFL with AND or OR gate. In C1-FFL with AND gate, the transcription of gene *Z* is dependent on the active form of both transcription factor *X* and *Y*. An inactive form of either transcription factor *X* or *Y* results in no transcription of gene *Z* and thus the production of gene *Z* product. In C1-FFL with OR gate, the transcription of gene *Z* is dependent on the active form of either transcription factor *X* or *Y*. Thus, an inactive form of either transcription factor *X* or *Y* does not affect the transcriptional process of gene *Z* and leads to the production of gene *Z* product.

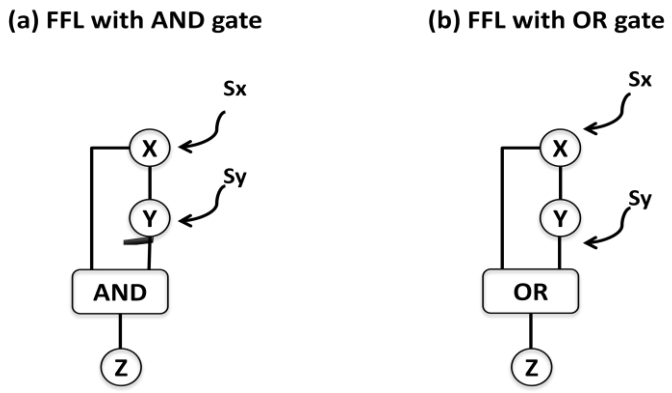

Figure 1.8: FFL with Boolean input gates: (a) FFL with AND gate; (b) FFL with OR gate (Alon, 2006).

1.6 Kinetics and dynamics of feed-forward loop

Combinatorial patterns of the FFL are easily enumerated, and their frequencies in nature have been empirically studied. The majority of FFL found in transcription networks of *E. coli* and *S. cerevisiae* were identified to be C1-FFL and I1-FFL (Mangan & Alon, 2003). In comparison to C1-FFL and I1-FFL, transcription networks belonging to other FFL types are relatively rare. In order to understand why the majority of these FFL are rare, it is necessary to understand their dynamics. Ingram *et al.,* (2006) suggested that the building structure of network motifs alone does not necessarily determine the biological functions without taking consideration of kinetic parameters, or dynamic time series experimental data.

Mathematical models have been widely applied to theoretically simulate the biological process such as *in silico* transcriptional regulation (Alon, 2006). Currently, the most common approaches used to investigate the kinetics or dynamics of FFL include

deterministic and stochastic mathematical models (Alon, 2006; Wilkinson, 2006). Such approaches predict the behaviour of an FFL by taking into consideration various factors or parameters such as promoter activity, activation and repression thresholds, and degradation/dilution rates of the gene products. Coherent type 1 FFL (C1-FFL) and incoherent type 1 FFL (I1-FFL) are FFL types have been intensively studied due to their high abundance and occurrence in the biological systems of both eukaryotes and prokaryotes (Mangan *et al.*, 2006; Mangan *et al.*, 2003). Consequently, the dynamics of these two FFL types are well understood. Here, we briefly review the well-known dynamics of these two FFL.

Figure 1.9 shows the dynamics of C1-FFL with AND gate with the presence (ON step) and absence (OFF step) of activation signal for transcription factor *X*. When the activation signal is for transcription factor *X*, S_x, is present, transcription factor *X* switches from inactive to active form and the concentration of activated transcription factor *X*, X^*, increases until it reaches its steady-state. When the concentration has increased to the activation threshold K_{xy}, transcription factor *Y* starts to be produced. It is then activated by the presence of its activation signal, S_y. The production of protein *Z* only begins when concentration of transcription factor *Y* reaches the activation threshold K_{yz}. The third gene, gene *Z*, is transcribed when both transcription factor *X* and *Y* are present. Thus, there is delay in the production of protein *Z* at the ON step because the production of protein *Z* requires activation of transcription factor *X* followed by production and activation of transcription factor *Y*.

When the activation signal S_x, is absent, transcription factor X switches from active to inactive form and the concentration of X^* decreases. When the concentration of X^* falls below K_{xy}, production of Z protein is switched off regardless of the presence of activated transcription factor Y. Thus, protein Z production is switched off without delay at the OFF step.

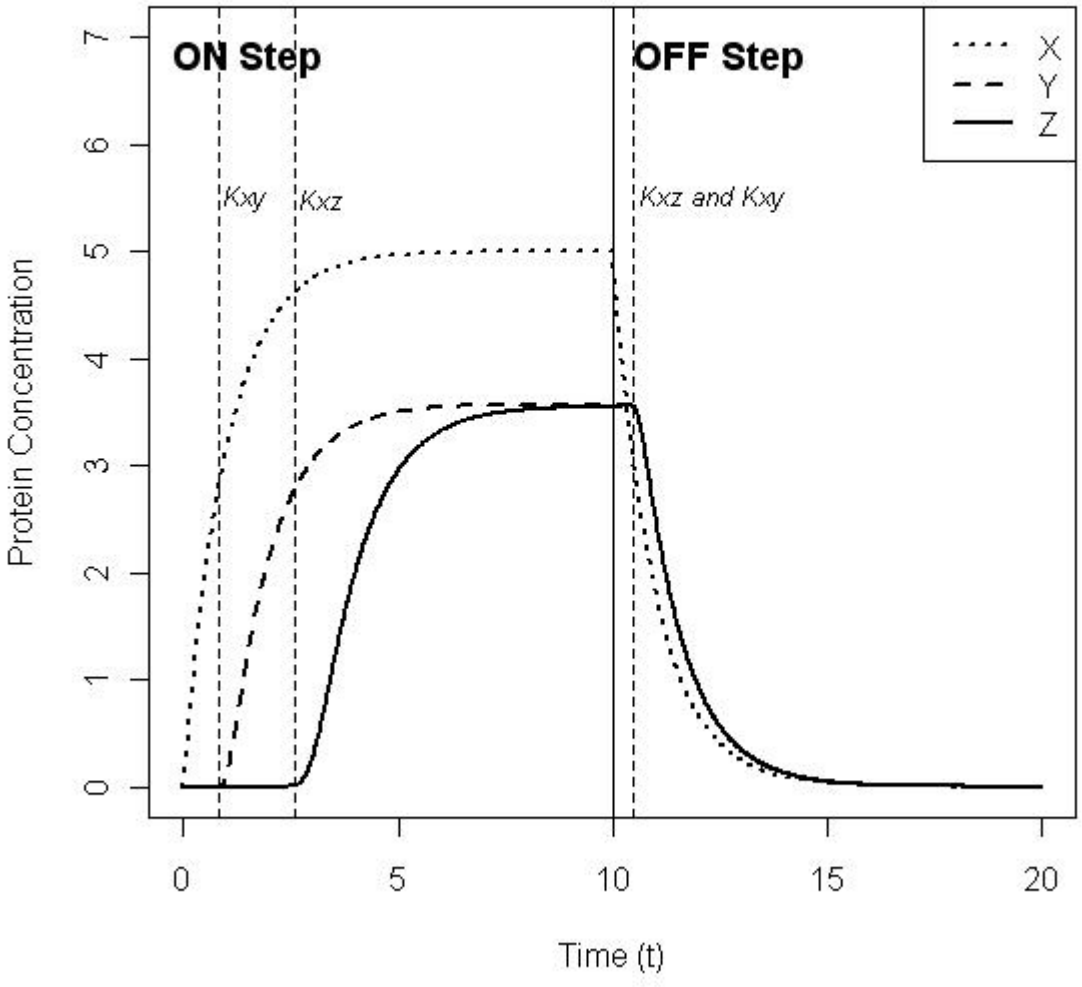

Figure 1.9: Dynamics of C1-FFL with AND gate. K_{xy} and K_{yz} represent activation thresholds for regulation of gene Y by transcription factor X and regulation of gene Z by transcription factor Y, respectively (Alon, 2006).

Figure 1.10 shows the dynamics of I1-FFL with AND gate in the presence (ON step) and absence (OFF step) of activation signal for transcription factor *X*. The dynamics of gene *Z* transcription is different from C2-FFL. There is acceleration in production of protein *Z* at certain time point when X^* reaches activation threshold K_{xz} and Y^* is present. The activation of protein *Z* production is enhanced and accelerated to the concentration higher than its steady-state. At the ON step, X^* exceeds K_{xy} and K_{xz}, and results in the cease of the production of protein *Y* and *Z*. Thus, there is no delay of repression of protein *Z* production.

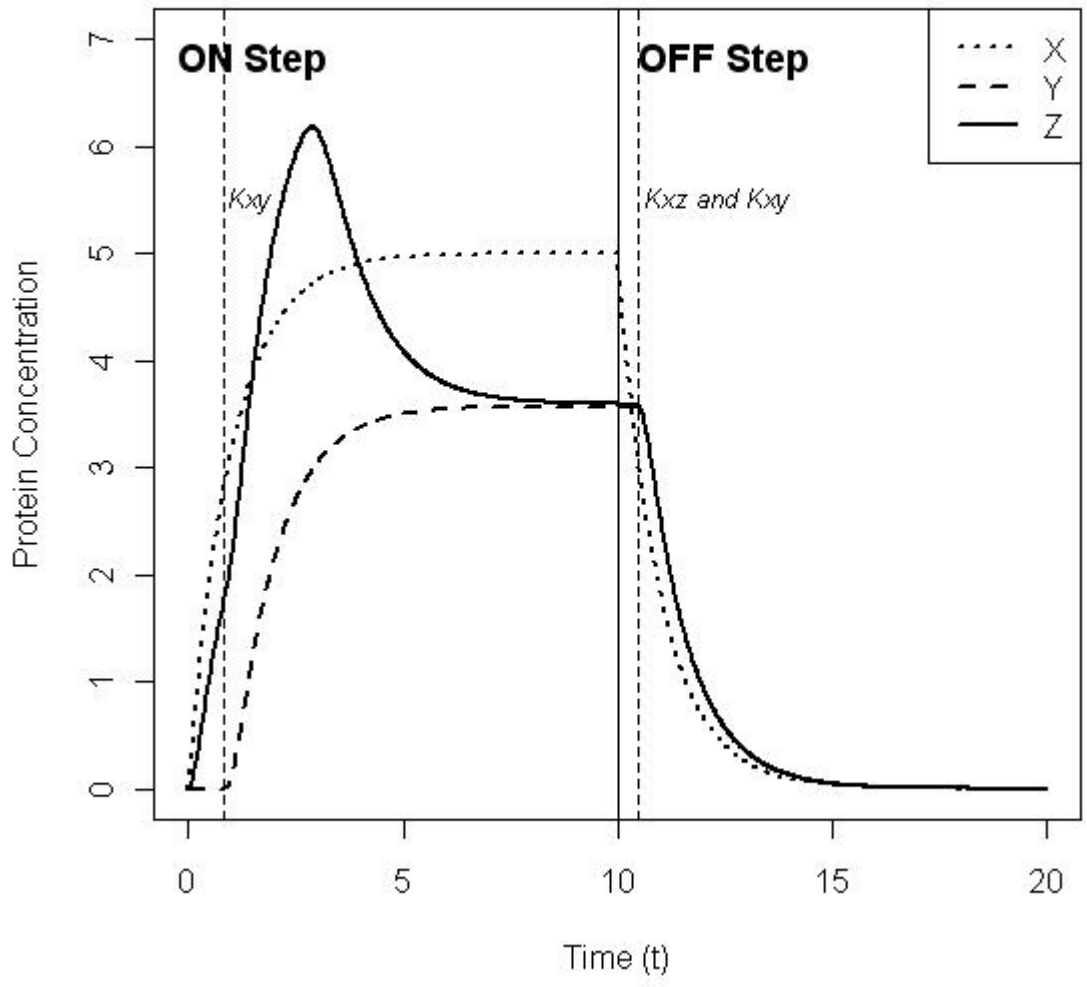

Figure 1.10: Dynamics of I1-FFL with AND gate. K_{xy} represents activation threshold and K_{yz} represents repression threshold for regulation of gene *Z* transcription by transcription factor *X* and transcriptional repressor *Y*, respectively (Alon, 2006).

1.7 Importance of FFL in biological systems

FFL has been identified in many important biological functions in both eukaryotic and prokaryotic organisms (Mangan *et al.*, 2003; Mangan *et al.*, 2006). Arabinose sugar utilisation in *E. coli* is an example of C1-FFL with AND gate logic. In this FFL, CRP

(Cyclic AMP Receptor Protein) is a universal transcription factor for sugar metabolism. In the presence of arabinose, activation of CRP by the activation signal cAMP (cyclic Adenosine Monophosphate) initiates transcription of *araC* gene. The product of *araC* gene is activated by its activation signal, L-arabinose. Both activated CRP and product of *araC* gene act as transcription factors to trigger transcription of the *ara* operon: *araBAD/araFGH* (Figure 1.11).

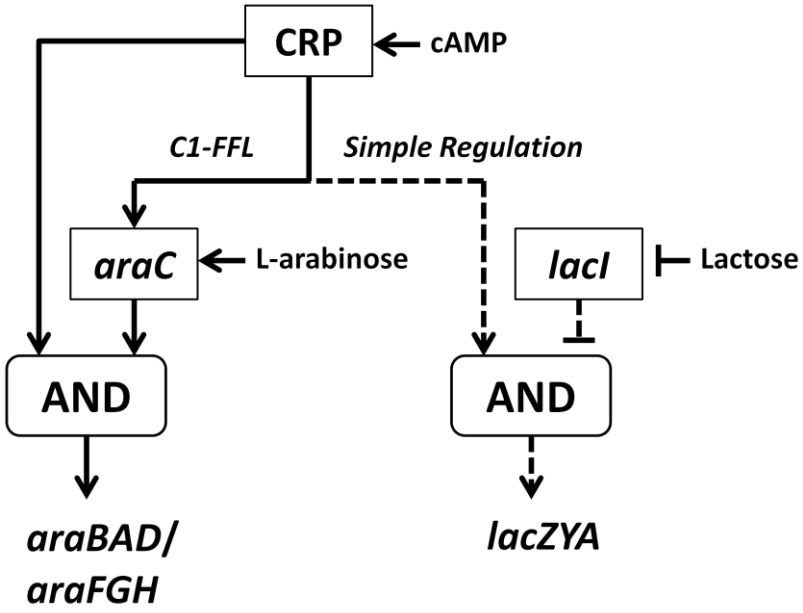

Figure 1.11: *ara* system (C1-FFL) and *lac* system (simple regulation) regulated by CRP activated by its activation signal, cAMP (Alon, 2006).

When lactose is present as sugar source, lactose metabolism is triggered. CRP is again activated by activation signal cAMP.The *lacI* gene, which encodes the inhibitor of *lac* operon, is repressed by lactose. With the presence of activate CRP and absence of LacI,

the transcription of *lac* operon: *lacZYA* is initiated. The whole transcriptional regulation is a type of simple regulation in which transcription of *lacI* gene is not regulated by CRP, in comparison to the transcriptional regulation of *ara* operon (Figure 1.11).

Glucose is the most favourite carbon source for *E. coli* (Desai & Rao, 2010; Nelson & Cox, 2000). Other sugars such as galactose, lactose, arabinose and xylose are used as carbon sources when glucose is depleted. Lactose is a disaccharide which is formed by galactose and glucose. Arabinose is a pentose consisting of aldehyde functional group. Due to the fact that lactose is partly formed by glucose, lactose is preferred over arabinose in *E. coli* as carbon source.

According to the dynamics simulation result discussed in Section 1.6, there is delay in *ara* enzyme production (C1-FFL with AND gate) and no delay in *lac* enzyme production (simple regulation). This is to ensure that *E. coli* will favour the utilisation of lactose in this case. Arabinose will be utilised only when lactose is depleted. FFL is also important in biological processes of eukaryotes. For examples, protein kinase Cα has involved in cancer development by promoting head and neck squamous cell carcinoma, via a feed-forward loop that leads to uncontrolled de-regulation in cell cycle (Cohen *et al.*, 2009).

1.8 Peculiar FFL types: The FFL types other than C1-FFL and I1-FFL

Figure 1.12 shows the distribution of FFL counts in transcription networks of *E. coli* and *S. cerevisiae* (Mangan & Alon, 2003). C1-FFL and I1-FFL make up the majority of the FFL network motifs. They are well studied in *E. coli* and *S. cerevisiae*. Mangan & Alon, (2003) found that 33 of 42 and 47 of 56 FFL in transcription network of *E. coli* and *S. cerevisiae*, respectively were identified to be C1-FFL and I1-FFL (Figure 1.12). Other FFL types are believed to have detrimental effects to the biological systems and were selected against during evolution. Ghosh *et al.* (2005) studied the noise characteristics of FFL and concluded that C1-FFL has the least noise among all the FFL.

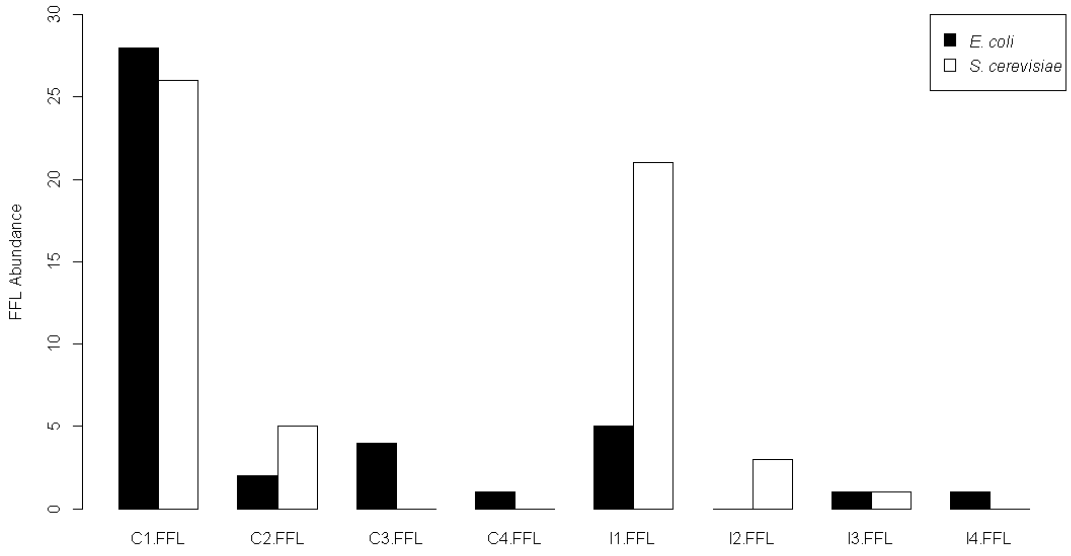

Figure 1.12: Distribution of FFL counts in transcription networks of *E. coli* and *S. cerevisiae* (Mangan & Alon, 2003).

Nevertheless, there are nine FFL in *E. coli* and nine FFL in *S. cerevisiae* belonging to peculiar FFL types which are present in low frequencies. In *E. coli*, these peculiar FFL participate in various physiological processes such as anaerobic respiration, maltose utilisation and osmoregulatory response (Chao *et al.*, 1997; Park *et al.*, 1997; Reidl & Boos, 1991; Unden & Bongaerts, 1997). These peculiar FFL types are believed to play important roles in biological regulation. Studying the dynamics of peculiar FFL using *E. coli*'s transcription networks as models will provide us with a theoretical basis to understand their function better and why they are not completely eliminated by natural selection.

1.9 Objectives of study

The present study aims to address two issues: (1) to simulate the dynamics of the peculiar FFL types: C2-FFL, C3-FFL, C4-FFL, I2-FFL, I3-FFL and I4-FFL with AND and OR gates at the ON and OFF step using deterministic approach; (2) to determine the advantageous or detrimental effects of peculiar FFL types on *E. coli* transcription networks.