Feed-Forward Loops in Eukaryotic Transcription Regulation

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

by

Wibowo Arindrarto

Graduate School of Life Science

Utrecht University

The Netherlands

September 2012

ABSTRACT

Cellular characteristics are determined by molecular interactions. These interactions are complex, forming multi-layered networks that influence each other. One way to analyze these networks is by analyzing their constituent motifs. There have been several transcription regulatory motifs thought to have been selected by evolution that confers specific traits to the cell, one example being the feed forward loop (FFL) motif. It is a motif formed by three distinct nodes consisting of two regulators and one target. Depending on the type of interaction between the nodes, the motif enables the target node to have distinct dynamics advantageous to the cell as a whole, making them interesting study subjects. These motifs were initially studied in simple unicellular organisms. Comprehensive motif studies on more complex organisms have been hampered by the lack of important data sets, notably the gene expression data and transcription factor binding data. FFLs have nevertheless been described in several small-scale studies in more complex eukaryotes. This thesis presents an overview of these small-scale studies by discussing their proposed FFL motifs and their relation to the initial proposed FFL motif in simpler organisms.

INTRODUCTION

Inside living cells, we find a myriad of chemical reactions take place between many different types of molecules. Together, these chemical reactions define a cell's behavior and its possible developmental trajectory. At the core of this multilayered network of chemical interactions, we find the transcriptional regulatory network. The transcriptional regulatory network is made up of interactions among transcription factor proteins and the genes that they regulate. The interactions in this network ultimately determine what proteins or other molecules is expressed at any given time. Thus, if we are to understand and control the behavior of a cell, we must understand the transcriptional regulatory network.

Such understanding is not an easy feat, however, as they are notoriously complex. For any given gene, several transcription factors could be involved in determining its expression level. The expression of these transcription factors themselves are in turn also regulated by other transcription factors, increasing their regulatory complexity. Moreover, many loops exist within the network, where the downstream products of a gene could be involved in its upstream regulation as well. Finally, for a given time or environmental condition, there could be different sets of regulators involved in determining a gene's expression level.

How then, can we start to understand the dynamics of a *transcription* network? Alon and colleagues proposed a reductionistic approach of analyzing smaller functional patterns that are found within the complete network[1]. They have discovered that some patterns, also called network motifs, occur more frequently than one would expect if the network in question were truly random. A random network, in this case, is defined as a network with a similar number of nodes but with its edges distributed randomly. One explanation behind the discrepancy of the motifs' occurrence in random versus real networks is that these motifs are products of natural selections. In other words, their apparent abundance is believed to confer one or more functional advantages that influence the cell's survival. Thus, understanding these motifs could help understand the behavior of network in which they are imbedded.

Alon's and colleagues approach is simple in essence, but quite revealing. A network motif consisting of n number of nodes (n=1, 2, 3, and so on) may be discovered by counting the occurrence of a network pattern consisting of the same number of nodes and comparing that number to the number it is expected to appear if the network is random. If a pattern occurs more times than expected, it is then deemed a recurrent network motif. This approach has been applied on a large-scale fashion to the whole known transcription network of the bacterium *Escherichia coli*[2] and the single-celled eukaryote *Saccharomyces cerevisiae*[3], yielding several network motifs with distinct characteristics. Importantly, these network motifs have proven useful in characterizing the behavior of the biological system they regulate.

A prime example is the bacterial flagella motor motif[4]. It is a simple motif consisting of three nodes: two transcription factors (FlhDC and FliA) and their target genes (*fli*LMNOPQR). Both FlhDC and FliA can activate the target genes, while FliA is also activated by FlhDC. Together they form a motif dubbed the type 1 feed-forward loop. As will be discussed later, the expression dynamics of *fli*LMNOPQR can indeed be predicted by the characteristics of this particular motif.

How far can these analyses be applied to other organisms? This is currently an open question, as our knowledge about gene regulation in more complex organisms is still very limited. Network motifs analyses require not only the knowledge about the interaction of transcription factors and their target genes, but also how much a presence or absence of a given transcription factor influences their target gene's expression levels. Additionally, given that organisms express a different set of genes in different environmental and/or developmental stages, these analyses are ideally performed in as many environmental and developmental stages as possible. All of this requires a vast amount of data to be generated.

Analyses on *E. coli* and yeast are already possible because of the wealth of transcription factor binding and gene expression data available on these organisms. They are arguably unmatched in the levels of gene expression data coverage compared to other organisms. There have been some attempts to characterize network motifs in complex organisms, such as human cells[5,6], but such studies is best considered preliminary at the moment given the sparseness of the underlying dataset. Finally, as a hint of how much information there is still left to discover, even the gene expression data for *E. coli* still receives continous updates[7,8].

Of course, that is not to say that the motifs found in *E. coli* and yeast do not exist in other cells. It is hard to resist thinking that such simple yet powerful motifs are not present in other organisms, conferring similar traits to the ones seen in *E. coli* and yeast. In this study, I aim to extend the analysis of network motifs to other organisms by drawing evidence from scientific literature. The focus will be specifically on a network motif called the feed-forward loop, a relatively simple motif that has been identified in *E. coli* and *yeast*. Given that no definitive studies on network motif identification have been performed on more complex organisms, the analyses will center on scientific literature discussing several gene circuits and how well these circuits fit the description of a feed-forward loop. Non-transcriptional networks comprising of RNA binding proteins and membrane proteins will also be discussed, to see whether it is possible to extrapolate findings from the transcriptional regulatory network into other networks if the topology is similar. Before going into these analyses, a brief introduction to the feed-forward loop motif is presented first.

THE FEED-FORWARD LOOP MOTIF

The feed-forward loop (FFL) motif is a simple network motif. It consists of three nodes: X, Y, Z, with X regulating the expression of Y and Z, and Y regulating the expression of Z (Figure 1)[1,9]. The regulators, X and Y, could either activate or repress transcription of their target genes, giving rise to a possible combination of eight different FFL motifs. These eight motifs can be categorized into two groups: coherent and incoherent loops, based on the coherence of input to the final product Z. In coherent loops, the input to Z that goes from both X and Y perform the same action, either activation (C1, C4) or repression (C2, C3). In incoherent loops, Z receives opposing signals from X and Y.

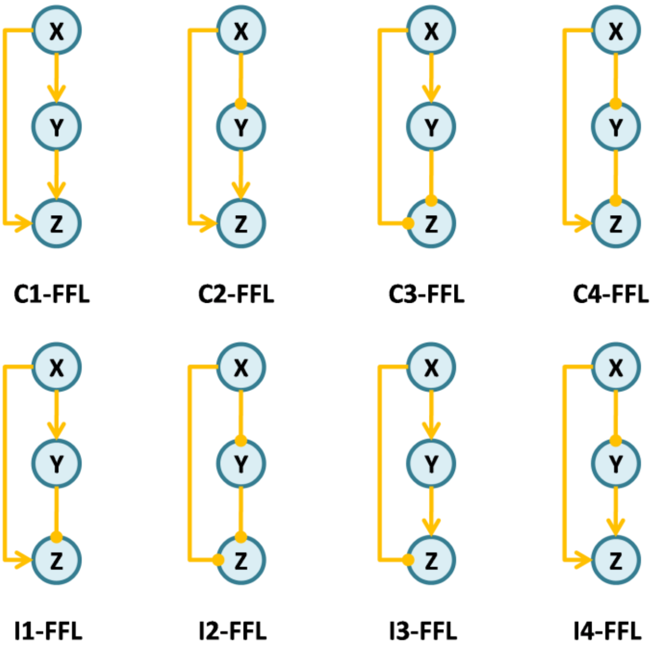

Figure 1 The eight possible feed-forward loop motif. Arrowheads denote activation, rounded ends denote repression. C: coherent, I: incoherent.

The utility of the network motif lies in its ability to predict the expression dynamics of the final target, Z. The prediction, however, has to take into account the kinetic parameters that govern the components' interactions. As will be shown below, the initial characterization done by Alon and colleagues assume a set of kinetic parameters that simplify modeling and calculation[9]. How far these assumptions always apply to all motifs remain to be shown.

Assuming X already reached its steady-state level, the expression level of Z (*dZ/dt*) then depends on its basal expression level (B_z), the concentration of Y and whether it is bound to its site or not (Y^*); whether it is available above the threshold level required to regulate Z expression (K_{yz}), and its decay rate (α_Z). The concentration of Y (*dY/dt*), similarly, depends on its basal expression level (B_y), whether X is activated and available above its decay rate (X^*), and its decay rate (α_Y). Expressed formally, these two terms are:

$$\frac{dY}{dt} = B_y + \beta_y f\left(X^*, K_{xy}\right) - \alpha_y Y$$

$$\frac{dZ}{dt} = B_z + \beta_z G(X^*, K_{xz}, Y^*, K_{yz}) - \alpha_Z Z$$

Where β_y and β_z are the expression level that will be integrated with the following functions f(u, K) or G(u_1, K_1, u_2, K_2). Additionally, the rate of X and Y activation is sometimes determined by another molecule, S_x or S_y (not shown in the formula). For example, X could be the phosphorylated by a kinase S_x before it can regulate transcription.

The first function depends on whether X is an activator or repressor, and the second function integrates the input of X and Y into Z, according to whether one transcription factor alone is sufficient to regulate Z (similar to an OR gate) or whether both transcription factors are needed (similar to an AND gate). Here, the output of the function is binary (yes or no), depending on whether the concentration of the active regulator (*X** or *Y**) is above or below the dissociation constant (K_{xy}, K_{xz}, or K_{yz}). In reality, however, transcription does not necessarily proceed in a binary fashion. Rather, its rate may change gradually as its regulators' concentration increase.

Still, using the above model, several prominent characteristics of the network motifs have been described. The C1-FFL, for example, has been shown to provide sign-sensitive delay response to the expression level of Z. Depending on whether the integration of X and Y into Z follows an AND or an OR gate, the expression level of Z may be delayed compared to when Z is simply regulated by X, without Y. If the integration, or sign, follows an AND gate, we see a delay during the rise of Z expression level. If the sign follows an OR gate, the delay is seen during the decay level of Z after the input X is deactivated. An example of the C1-FFL with an AND gate is the arabinose system in *E. coli*, where arabinose producing enzymes are produced only when the CRP protein is active and the protein *Ara*C is present above the threshold[10]. The bacterial flagella motor, as mentioned earlier, is an example of a C1-FFL with an OR gate function.

An example of the incoherent loops is the I1-FFL. The network motif has been shown to generate pulses of transcription levels, in which the initial expression level is high and then reduced after a certain period. The initial high expression is because the Z protein is expressed due to X's activating action, but after Y reaches its threshold for regulating Z, Y acts as a repressor that reduces Z expression to its steady state. An example of this motif is the galactose system in *E. coli*, where the initial production of the galactose metabolizing enzyme is high but then toned down[11,12].

As mentioned earlier, the FFL characterization in *E. coli* and yeast benefit from the the presence of many transcription factor binding and gene expression datasets. In more complex organisms, such interaction databases are not yet sufficiently comprehensive, limiting the study of network motifs. However, there have been several published studies which investigated individual FFL motifs in different multicellular animals and tried to characterize their properties in relation to the biological phenomenon they regulate. Here, these examples are discussed in detail, beginning with the discussion of network motifs composed of transcription factor proteins (hereafter termed canonical FFL).

CANONICAL FEED-FORWARD LOOP MOTIFS

The first example of a known FFL discussed in eukaryotes is the FFL responsible for mouse skeletal muscle differentiation[13]. Here, the components involved are the transcription factors MyoD, Mef2D, several muscle-specific genes, and the protein kinase p38 (Figure 2). The p38 protein may seem oddly placed at first, but it is really the S_y component of the motif that was also described in the previous section. Its role is to activate the Y node, in this case Mef2D, and may be considered part of the network motif as well.

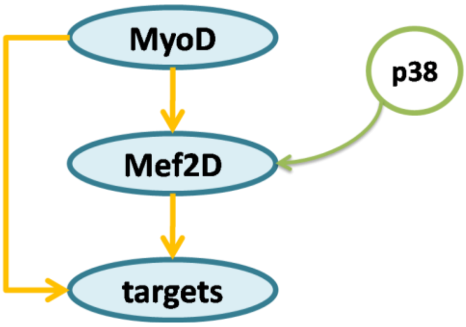

Figure 2 The FFL consisting of MyoD, Mef2D, p38, and their target genes. Orange arrows denote transcriptional regulation, green arrow denotes phosphorylation.

Before looking into the FFL, we first look into the biological process it is involved in regulating. MyoD and Mef2D are two transcription factors regulating skeletal muscle differentiation. The differentiation process occurs in several stages that can be broadly divided into two: the early stage and the late stage[14]. In the early stage, myoblasts start to express muscle-specific cells under the control of several transcription factors, MyoD and MEF2D among them. These cells then become a differentiated cell type called myocytes. Several myocytes then fuse together to form multinucleated myotubes in the late differentiation stage.

Several lines of evidences support the hypothesis that MyoD, Mef2D, and p38 forms a FFL to regulate the expression of muscle specific genes[13]. First is the physical interaction of the motifs' nodes. Using chromatin immunoprecipitation, MyoD and Mef2D have been shown to bind and activate the muscle specific genes[14]. These target genes are only expressed when MyoD is present in its activated form. Interestingly, the timing of this binding is influenced by the activation of p38. Earlier p38 activation was shown to result in an earlier binding of Mef2D. This suggests that MyoD, Mef2D, and p38 form a network that results in a multi-step activation of its target genes.

The second line of evidence is the expression and activity of the three nodes. Mef2D is a known phosphorylation target of p38. When p38 phosphorylation is inhibited, the expression of late-stage, muscle-specific genes are inhibited as well. Conversely, precocious overexpression of p38 resulted in the precocious expression of Mef2D and its target genes. When this precocious expression of p38 is combined with early expression of Mef2D, an even earlier expression of the late-stage genes are observed. However, early expression of Mef2D without p38 did not result in a similar precocious expression of early target genes, suggesting that Mef2D needs to be phosphorylated before it can activate its targets. Finally, both p38 and Mef2D seems to be dependent on MyoD expression for their own expression.

Taken together, this suggests that the transcription factor MyoD, Mef2D, and the protein kinase p38 is involved in an FFL. The proposed motif can be further grouped into the coherent type 1 FFL based on each element's interaction with each other. This particular motif has a characteristic of introducing delay in the expression or repression of the final product. The delay is present because Mef2D needs time to be expressed and activated. If the time it takes for Mef2D to be activated is shortened, the target genes will also be expressed sooner. Indeed, this is what was shown in the study.

However, even though the authors reach the conclusion that MyoD, Mef2D, and the target genes are involved in a type 1 FFL, it was never directly shown that MyoD binds directly to the promoter of Mef2D even though Mef2D expression increases when MyoD is active. How then, can the behavior of the network be similar to a type 1 FFL? As mentioned earlier, the behavior of the network's final node depends not only on the network topology but also on the reaction kinetics. According to Alon's model, the long path of the FFL (the path that goes from MyoD, Mef2D, and the targets) increases the amount of time it takes for the final target genes to be affected by Mef2D. Thus, adding an extra node to this path essentially increases its delay. In other words, MyoD may not have to interact directly with the Mef2D promoter in order for the network motif to show characteristic type 1 FFL behavior.

Finally, it is worth considering what benefits the type 1 FFL motif here endows the cell or organism as a whole. As mentioned earlier, network motifs are thought to confer one or several advantages since they are selected by natural selection. For the MyoD motif, the clue might be found in an earlier experiment. It was found that different clusters of target genes could be made according to the timing of expression relative to MyoD. The early gene cluster has a higher representation of nuclear regulatory factors and genes involved in extracellular matrix processing. The late gene clusters, on the other hand, have a higher representation of structural and cytoskeletal genes specific to skeletal muscle cells. Delaying the expression of genes in the late cluster is a possible mechanism for the cell to allow genes expressed in the early cluster to complete their function.

The next motif that will be analyzed is involved in the neuronal subtype specification network in the fruit fly *Drosophila melanogaster*[15]. In this case, several FFLs are involved, creating an elegant cascade of temporal gene expression that regulates cell differentiation. The approaches employed by the authors are different from the previous motif discussed; using knockout mutants and immunostaining *in vivo*. This approach has the advantage of eliminating in vitro artifacts but comes at the cost of having reduced quantitative power. These approaches can also be used to infer the underlying network topology, albeit with some drawbacks.

Before examining the network, it is useful to consider the biological phenomenon in question. In *D. melanogaster*, the ventral nerve cord of the developing embryo contains a lateral cluster of four neurons that develop from a single neuroblast[15]. There are two cells in the four neural cluster that produce neuropeptides: the Ap1/Nplp1 cell, producing the FMRFa neuropeptide and the Ap1/FMRFa cell that produces the Nplp1 neuropeptide. The other two cells are called Ap2 and Ap3 and are considered similar. Thus, from a single neuroblast, four cells of three different types are formed. How were these cells specified?

A combination of knockout mutants and immunofluorescence allows the analysis of transcription factors known to be involved in the neuronal specification by counting the number of cells expressing a given marker in a knockout mutant. It was found that three different FFL are involved in the cell type specification and that the specification events hinge on the characteristic of the apparent feed-forward loop motifs (Figure 3). The first network specifies the Ap1/Nplp1 cell, the second network suppresses the first network and specifies the Ap2/3 cells, while the third one specifies the Ap4/FMRFa cell. From the initial asymmetric division of the neuroblast, the resulting daughter cell express the *col* gene, which will activate the FFL required for the Ap1/Nplp1 cell specification. The gene upstream of *col*, on the other hand, also activates *sqz* and *nab*, forming another FFL. Both *sqz* and *nab* can repress *col*, but this do not happen in the first daughter cell of the neuroblast because *nab* is not expressed as early as *col*. When *nab* does get expressed, it suppresses *col* and results in the specification of the Ap2/3 cells, which emerge later than the Ap1/Nplp1 cell. Finally, the Ap4/FMRFa cell is specified by the action of *grh* that inhibits *cas* expression and activates other targets required for Ap4/FMRFa fate.

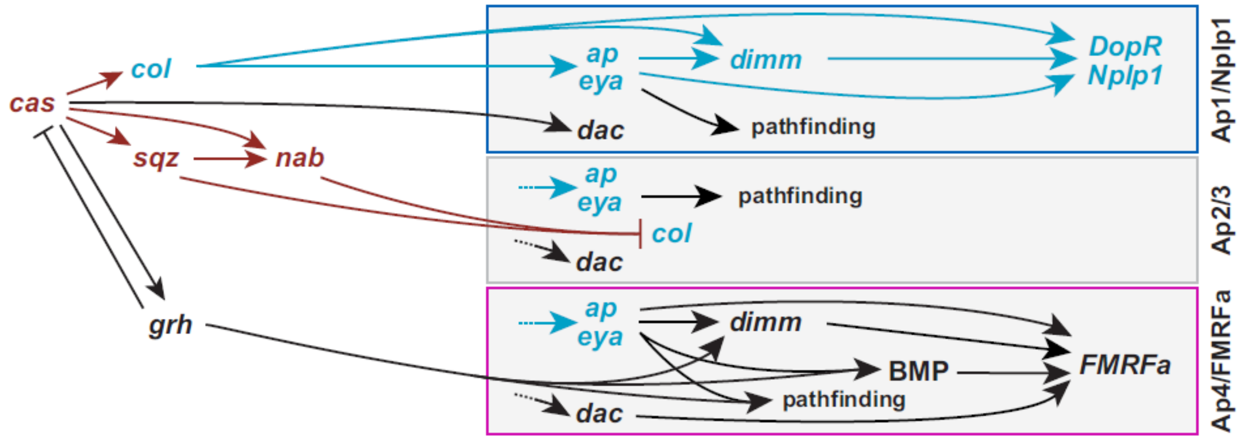

Figure 3 The FFLs the specify the D. melanogaster Ap neuron cluster.

At a glance, each of these FFL seems to be similar to the FFL described in yeast and *E. coli*. However, the proposed model here was made based on the expression of the genes without checking if any physical contact between the transcription factors and their target genes' regulatory region occurs. Without taking this into account, the full extent of the network remains to be seen since there might be other intermediate factors that exist between a transcription factor expression and the expression of their targets. Indeed, it seems rather odd that the final FFL that specifies the Ap4/FMRFa are activated later than the second FFL, even though the author indicates a direct regulation of the FFL through *grh*. There could be other factors or even other FFL motifs that delay the expression of *grh*. Nevertheless, their argument on the involvement of two FFLs on the specifications of Ap1/Nplp1, Ap2, and Ap3 stands. Their proposed network possesses similar characteristics to the FFL, which in this case is the C1-FFL.

The final example of this section provides another view on how FFLs can be utilized for cell differentiation. Similar to the neuronal subtype specification, the animal is *Drosophila* and the biological process involves different FFLs that are interconnected together. However, different from the neuronal subtype specification, this example shows how the incoherent FFL can also be used to specify cells. The biological phenomenon in this example is the differentiation of the photoreceptors in the *Drosophila* eye[16].

During the course of embryonic development, the animal's eyes are specificied from a single layer of epithelium cells. The process that has been known to utilize transcription factors arranged in multiple FFL is the specification of photoreceptors. The fully developed eyes of *Drosophila* are actually composed of about 800 single unit eyes, called ommatidia. A single ommatidium, is composed of eight photoreceptors (PRs) that can be grouped into six outer PRs (R1-6) and two inner PRs (R7-8). The outer PRs are distinguished by their expression of the light-sensitive protein Rhodopsin 1 (*Rh1*) which is mainly used for motion detection. On the basis of Rhodopsin expression, the inner PRs are also further divided into two groups: the "pale" (p) group and the "yellow" (y) group. It is this grouping that determines the subtype of the whole ommatidia. Pale group inner cells (pR7 and pR8) express the Rhodopsin Rh3 and Rh5 respectively, while the yellow group inner cells (yR7 and yR8) express Rh4 and Rh6 respectively. Both Rh3 and Rh4 are ultraviolet-sensitive Rhodopsins while Rh5 and Rh6 each detect a different range of the visible light wave spectrum. The location of these cell subtypes are random, although they both occur at a relatively fixed proportion, 35:65 for pale:yellow.

Using a series of knockout mutants and gel-shift assays, it has been proposed that the specification of a subset of these photoreceptors is controlled by series of interconnected FFL (Figure 4). The subset in question includes the outer PRs and the inner pale R7. Here, the two regulators are the transcription factors Orthodenticle (*Otd*) and Defective proventriculus (*Dve*), since they are both present in all the FFLs that govern the expression of various Rhodopsins in different PRs. *Otd* on its own could activate the expression of Dve and Rhodopsins *Rh3*, *Rh5*, and *Rh6*. *Dve*, on the other hand acts as a repressor for these Rhodopsins. Thus, we see that *Otd* and *Dve* forms an incoherent FFL type 1 (I1-FFL), given their different action on the Rhodopsins. This is indeed what was observed in the R1 to R6 cells, where *Rh3*, *Rh5*, and *Rh6* are all repressed.

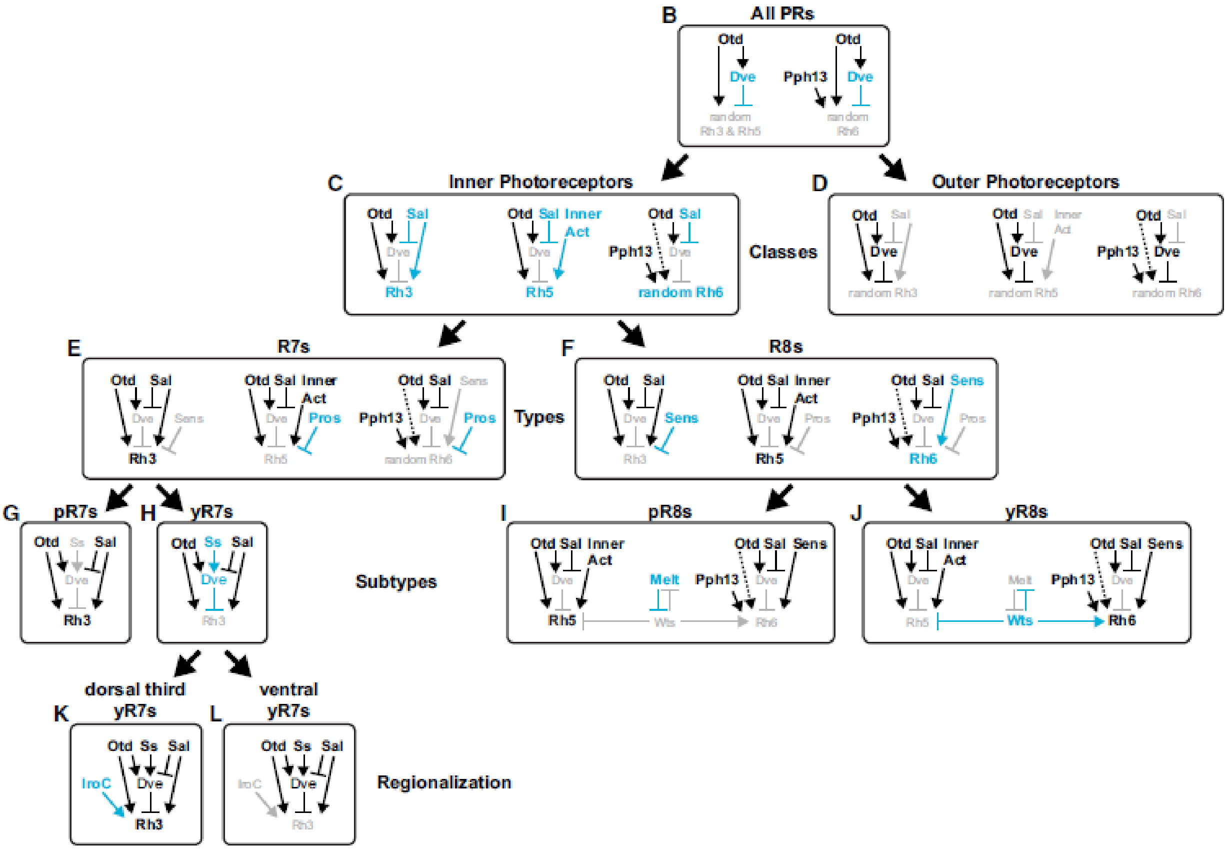

Figure 4 Scheme of gene regulatory networks involved in *Drosophila* ommatidia subtype specification. The FFLs are present in boxes B, C, D. For each level, blue color indicates a new component compared to the previous level.

In the PR7 cells, the repression of *Rh3* by Dve is alleviated by the presence of another transcription factor called *Sal*. *Sal* can also function as an activator of *Rh3*, thus effectively overriding the initial I1-FFL with a new network motif. The new motif, consisting of *Sal*, *Dve*, and *Rh3*, is a coherent type 4 FFL, given the similar outcome of *Rh3* expression by the action of *Sal*. However, given that *Dve* initially also represses *Rh5* and *Rh6*, one might wonder how pR7 cells maintain repression of Rhodopsins when *Dve* is not expressed anymore. Here, it was found out that the cells actually express another repressor called *Pros*, that is able to repress *Rh5* and *Rh6* without affecting *Rh3* expression. In the other PRs, a similar logic of overriding the initial I4-FFL to allow for *Rh5* or *Rh6* expression is also present. However, they do not form new FFLs since the regulators that repress *Dve* and activate the Rhodopsins have not been shown to be the same protein.

EXTENDED FEED-FORWARD LOOP MOTIFS IN EUKARYOTES

After looking at several studies on regular FFLs, which are FFLs whose X and Y nodes are transcription factors, several 'extended' FFLs are presented. Extended here means that one or more components of the FFL is not a transcription factor. They could be membrane proteins, RNA-binding proteins, or some other proteins. Although these FFLs are not FFLs in the proposed definition, it will be shown that they in fact behave similarly to regular FFLs.

The first example discussed is a simple motif found in yeast consisting of the RNA-binding protein KHD1, the transcription factor ASH1, and the cell wall protein FLO11[17] (Figure 6). This loop controls the transition of yeast cells into its filamentous form. In yeast cells, the cells can reversibly differentiate into a filamentous form when they are placed under nitrogen stress. In this FFL, the first component of the FFL is KHD1, which is able to repress FLO11 and ASH1 translation. ASH1 itself can activate the transcription of FLO11. It is apparent that this FFL is of the coherent type 2 subtype (C2-FFL), given the arrangements of the components.

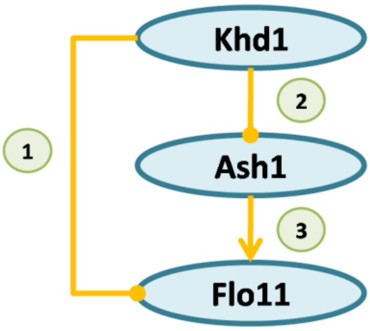

Figure 5 The FFL consisting of the RNA-binding protein Khd1, transcription factor Ash1, and cell wall protein Flo11.

How was this network motif deduced? The interaction between *khd1* and *flo11* (Figure 6, arrow 1) was deduced by knocking out *khd1* and observing that *flo11* mRNA levels increases as a result. The second interaction, between KHD1 and *ash1* mRNA, was deduced by using CLIP (cross-linking immunoprecipitation). Using this method, the authors of the study found the binding motif of KHD1 and showed using a GFP construct that the motif can be used to repress expression. Finally, the last interaction was uncovered by knocking out *ash1*, which resulted in a drop of *flo11* mRNA levels.

How similar is this network to the regular FFL consisting of only transcription factors? The regular C2-FFL model[9] predicts that the final product, FLO11, should have an initial delay in expression after the activity of the first component, KHD1, is gradually lost. Unfortunately, no such measurement was made in the study. Instead, the authors postulated that the FFL topology here was present to ensure that transitions between the filamentous form to non-filamentous cells is bistable; either the yeast becomes filamentous or not. In the FFL motif, it is apparent that KHD1 regulates the translation of FLO11 directly and its transcription indirectly through ASH1. When a mother cell detects that it should produce progenies with the regular cellular form, then expressing KHD1 would ensure that FLO11 is not expressed anymore. In this case, the initial condition is different from the one described in the regular C2-FFL model since at the steady state, KHD1 is not present. It is still possible that during the switch from regular cell morphology to filamentous form there is an initial delay as predicted by the model. However, further experimental data is required to confirm this.

Finally, it should be noted that some of the interactions shown in this model may not be direct physical interactions as well. The only physical interaction suggested by the experiments was that of KHD1 and *ash1* mRNA. The interactions between KHD1 and *flo11* and/or *ash1* and *flo11* may have other intermediaries. Although this does not prevent the motif from behaving like a canonical FFL, as has been shown in previous examples, it does not conform to the initial FFL motif description.

The second extended FFL that will be discussed is present in the social Amoeba *Dictyostelium discoideum*[18]. The social Amoeba is a model organism that has been commonly used to study morphogenesis and cell differentiation. Though able to live its entire life as a single-celled organism, *D. discoideum* can aggregate to form multicellular structures when environmental conditions are adverse. This multicellular structure then undergoes several steps of morphogenesis until it forms a fruiting body that is able to disperse aerosol spores, sending its progeny to live in another place.

A feed-forward loop motif has been indicated to regulate the initial differentiation event that takes place after *D. discoideum* cells aggregate[19]. It consists of the proteins GBF, a transcription factor, and LagC, a membrane protein, both regulating the expression of 16 other genes (Figure 2). Several lines of evidence point to the conclusion that GBF and LagC are part of an FFL. The initial evidence is the difference of expression times. GBF mRNA is expressed after 4 hours of aggregation stimulation, while LagC is expressed after 8 hours. The second line of evidence comes from the knockout mutant expression patterns. LagC is not expressed when GBF was knocked out, but GBF is still expressed even though LagC is knocked out, indicating that GBF is required for LagC expression but not vice versa (Figure 7, arrow 2). Furthermore, both GBF and LagC are required for the expression of the 16 target genes. When LagC is knocked-out, these 16 genes are not expressed (Figure 7, arrow 3). Similarly, when GBF is knocked-out and LagC expressed under a GBF-independent regulatory region, the target genes are not expressed (Figure 7, arrow 1). Overexpression of either LagC or GBF when the other is knocked-out also did not result in the expression of their target genes. The target genes are only expressed when both GBF and LagC are expressed, suggesting that they both regulate the target genes. Finally, the 16 target genes of GBF and LagC are expressed after both proteins are expressed, in line with the idea that GBF and LagC form an FFL.

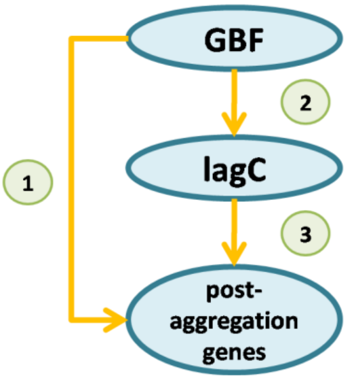

Figure 6 The FFL of *D. discoideum* that controls the expression of post-aggregation genes[19].

In the case of GBF and LagC, the FFL in question seems to be the Type 1 Coherent FFL (C1-FFL). The property of this type of FFL is that it introduces delay in transcription, depending on its logic. The fact that both GBF and LagC is required for the expression of the target genes shows that their operational logic is similar to the

AND gate. Indeed, the delay of transcription is apparent in the study that was done. This is useful to the organism because it needs to be sure that genes required for its post-aggregation phase are expressed after aggregation has occurred.

Going back to the motif, it is apparent that there are similarities between the regular and extended FFL. How is this achieved? Looking closer at this network's components, it is clear that LagC is a membrane protein. This necessitates the presence of downstream signaling molecules that in turn will interact with the regulatory region of the target genes to activate its expression. In essence, this is still similar to the path from the first protein (protein X) to the final component (protein Z) through protein Y. In the regular FFL, it is that path that introduces delay of expression in the first place, given the model's assumption. Here, the path from LagC to its target molecules is basically extending the number of steps already present between GBF to its target genes. One can assume that the presence of signaling from LagC then increases the delay time similar to a regular FFL.

LARGE SCALE NETWORK MOTIF STUDIES

The examples given so far represent only a minute subset of the possible FFL interactions in eukaryotes. However, it is apparent from these examples that understanding simple network motifs is useful for understanding or predicting the behavior of a larger system. There have been attempts to characterize eukaryotic network motifs in a more comprehensive manner. One study investigated the regulatory circuitry formed by the transcription factors Oct4, Sox2, and Nanog in human embryonic stem cells[20]. Using chromatin immunoprecipitation and microarray analysis, the study found two distinct motifs: the autoregulatory loop, where the product of a gene regulates its own expression, and the feed-forward loop. Approximately 353 protein coding genes and 2 miRNA were found to be part of these motifs. Another study looked at the network motifs that might exist in human hepatocytes with similar approaches[21]. Here, six transcription factors with known roles in hepatocyte biology were investigated. Similar to the network in human embryonic stem cells, this study also found the feed-forward loop network motifs with at least 246 distinct genes involved. While these studies only investigated a small subset of the network in the respective model cells, the fact they identified the FFL motif suggest that the motif is indeed important for higher organisms as well.

One important aspect to remember regarding these high-throughput, large-scale studies is that they only investigate a portion of the network motifs' characteristics. In the case of the two previous studies, the conclusion was obtained from transcription factor binding data and large-scale gene expression measurement. In order to characterize the network better, the specific biological processes regulated by these motifs should also be characterized so the relation between the network motif and the actual biological process can be established. Until now, the current eight FFL subtypes' characteristics are made with several assumptions. More experimental data to validate these predicted behavior would only be needed.

Finally, a critical point to remember is that these smaller and simple networks in reality are inseparable from the larger networks. In order to make sense of the larger network, one needs to integrate the smaller network motifs. This integration itself might introduce problems when conflicting predictions are found. An example is presented here, based on the data obtained by Alon and colleagues[9] (Figure 5).

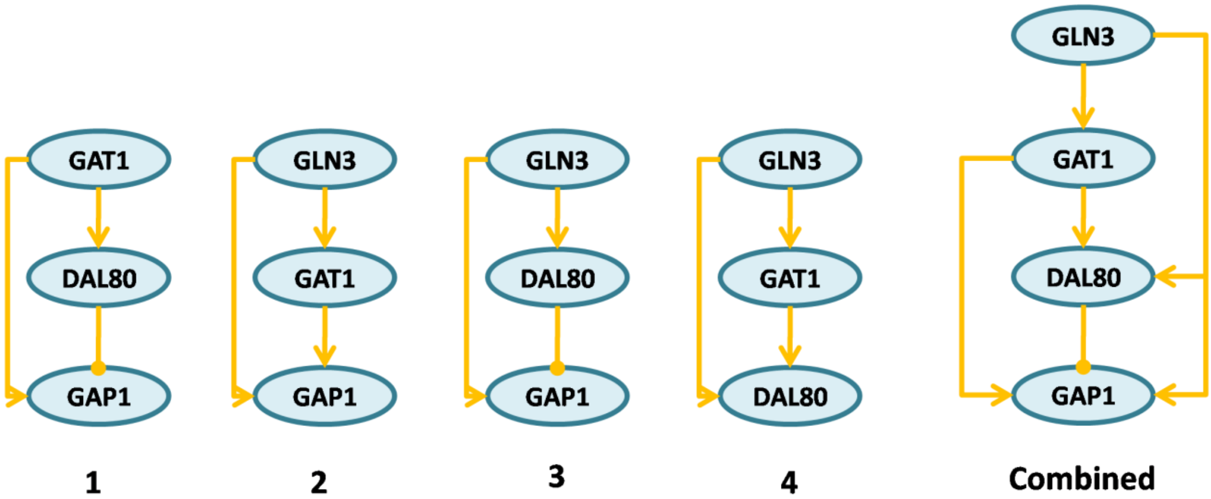

Figure 7 Integration of separate FFLs into the larger circuit. Notice that GAP1 has two different input types: repression from DAL80 and activation from GAT1 and GLN3.

Here, four FFL that are actually part of the same bigger network are presented. Two of these FFLs are of the C1-FFL subtype (2 and 4), while the others are of the I1-FFL (1 and 3). Considered separately, one might be able to predict the behavior of the final component (Z) of each network. However, when combined into the larger network, a problem arises. In the larger network, it is unclear how GAP1 behaves: does it have a delayed expression timing, as predicted by the C1-FFL, or is it expressed in a pulse-like manner, as predicted by the I1-FFL? The answer can be obtained if we also consider the kinetic parameters of the real interaction. In reality, GAP1 has to behave in a certain way. Our current FFL models assume that interaction between a motif's components is either present or not. These assumptions do not capture the natural condition, where instead of a present/not condition we find a more gradual condition.

CONCLUSIONS

The examples of FFL presented here are far from exhaustive. Despite that, some general trends may be derived. In eukaryotic cells, it seems that regular FFL are employed during cellular differentiation processes that involve multiple steps or temporal control. Here, the most frequent FFL is the coherent FFL, since it is the subtype capable of introducing transcription time delays and thus imbue the network with temporal control. It is unclear how far the other FFL subtypes may play a role. This might be caused by publication bias, as most published FFLs seem to be coherent FFLs.

Large-scale studies, where the network motifs are determined based on protein binding and expression data, needs complementary data on the interaction kinetics if we are to characterize the resulting motifs. This is because integration of the motifs alone could lead to conflicting conclusions given the assumptions in the current model. Thus, quantitative interaction kinetics are important in understanding the data as well.

In the extended FFL example, it was shown that sometimes the networks that contain non-transcription factor proteins can also behave similarly to regular FFL. This hints to a possible use of network motifs in analyzing different molecules that have similar interaction characteristics.

REMARKS AND FUTURE DIRECTIONS

Network motifs analysis is a relatively recent development in the field of biology. Its emergence was a response to increasingly complex biological interactions that we have only begun to appreciate. In a way, it embodies the classical reductionist paradigm, where it is believed that one can deduce general characteristics of a complex system by analyzing representative parts of it. Indeed, the approach has shown its usefulness in analyzing several transcription regulation systems in simple organisms and its utility is expected to increase as various transcription data sets for more complex organisms are created.

Initial analysis in the early 2000s focused on identification of important motif and their characterization. Such analyses, as mentioned earlier, depend on a well-defined interaction datasets. For organisms more complex than *E. coli* or yeast, we have yet to gather enough datasets. Thus, it is clear that the path forward must involve the collection of more datasets, particularly gene expression and transcription factor binding datasets, under as many conditions as possible.

Moreover, it is also important that correct terminology is applied during this data collection phase. For example, the author found that many papers claiming to have found a feed forward loop motif are not entirely correct. If we take the initial study from Alon and colleagues as a reference, then by definition an FFL should only contain three nodes with direct interaction among them that goes in one direction. Other authors frequently use the term loosely, claiming that a certain motif is an FFL even though a feedback loop is clearly present. Other authors claim their motif is an FFL even though it contains many more than three nodes with more complex interaction as well. The latter could behave similarly to an FFL, as shown in some early examples in this thesis. However, it still deviates from the original FFL description given by Alon and colleagues.

The study of network motifs is still in its infancy. It is poised to grow, given the increasing ease and decreasing costs of performing high throughput experiments. It is likely that we will discover more functionally important motifs in the future. With the advance of analyses methods, we may also be able to discover more complex motifs with their own unique characteristics.

References

1. Alon, U. Network motifs: theory and experimental approaches. *Nature reviews. Genetics* **8**, 450–61 (2007).

2. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics* **31**, 64–8 (2002).

3. Lee, T. I. *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science (New York, N.Y.)* **298**, 799–804 (2002).

4. Kalir, S., Mangan, S. & Alon, U. A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli. *Molecular systems biology* **1**, 2005.0006 (2005).

5. Kim, T., Kim, J., Heslop-harrison, P. & Cho, K. Evolutionary design principles and functional characteristics based on kingdom-specific network motifs. *Bioinformatics* **27**, 245–251 (2011).

6. Tsang, J., Zhu, J. & van Oudenaarden, A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Molecular cell* **26**, 753–67 (2007).

7. Huerta, A. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Research* **26**, 55–59 (1998).

8. Gama-Castro, S. *et al.* RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic acids research* **39**, D98–105 (2011).

9. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11980–5 (2003).

10. Mangan, S., Zaslaver, a & Alon, U. The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks. *Journal of Molecular Biology* **334**, 197–204 (2003).

11. Kaplan, S., Bren, A., Dekel, E. & Alon, U. The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Molecular systems biology* **4**, 203 (2008).

12. Kuttykrishnan, S., Sabina, J., Langton, L. L., Johnston, M. & Brent, M. R. A quantitative model of glucose signaling in yeast reveals an incoherent feed forward loop leading to a specific, transient pulse of transcription. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16743–8 (2010).

13. Penn, B. H., Bergstrom, D. A., Dilworth, F. J., Bengal, E. & Tapscott, S. J. A MyoD-generated feed-forward circuit temporally patterns gene expression during skeletal muscle differentiation service A MyoD-generated feed-forward circuit temporally patterns gene expression during skeletal muscle differentiation. *Genes & Development* 2348–2353 (2004).doi:10.1101/gad.1234304

14. Lluís, F., Perdiguero, E., Nebreda, A. R. & Muñoz-Cánoves, P. Regulation of skeletal muscle gene expression by p38 MAP kinases. *Trends in cell biology* **16**, 36–44 (2006).

15. Baumgardt, M., Karlsson, D., Terriente, J., Díaz-Benjumea, F. J. & Thor, S. Neuronal subtype specification within a lineage by opposing temporal feed-forward loops. *Cell* **139**, 969–82 (2009).

16. Johnston, R. J. *et al.* Interlocked feedforward loops control cell-type-specific rhodopsin expression in the Drosophila eye. *Cell* **145**, 956–68 (2011).

17. Wolf, J. J. *et al.* Feed-forward regulation of a cell fate determinant by an RNA-binding protein generates asymmetry in yeast. *Genetics* **185**, 513–22 (2010).

18. Chisholm, R. L. & Firtel, R. a Insights into morphogenesis from a simple developmental system. *Nature reviews. Molecular cell biology* **5**, 531–41 (2004).

19. Iranfar, N., Fuller, D. & Loomis, W. F. Transcriptional regulation of post-aggregation genes in Dictyostelium by a feed-forward loop involving GBF and LagC. *Developmental Biology* **290**, 460 – 469 (2006).

20. Boyer, L. a *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–56 (2005).

21. Odom, D. T. *et al.* Core transcriptional regulatory circuitry in human hepatocytes. *Molecular Systems Biology* 1–5 (2006).doi:10.1038/msb4100059